Medical Image Analysis 67 (2021) 101816

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Selective synthetic augmentation with HistoGAN for improved histopathology image classification

Yuan Xue[a,1], Jiarong Ye[a,1], Qianying Zhou[a], L. Rodney Long[b], Sameer Antani[b], Zhiyun Xue[b], Carl Cornwell[b], Richard Zaino[c], Keith C. Cheng[c], Xiaolei Huang[a,*]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA
[b] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD 20892, USA
[c] Department of Pathology, Penn State Health Milton S. Hershey Medical Center and Penn State College Of Medicine, Hershey, PA 17033, USA

ARTICLE INFO

Article history:
Received 2 March 2020
Revised 24 July 2020
Accepted 14 August 2020
Available online 1 October 2020

MSC:
68T05
68T45

Keywords:
Histopathology image classification
Medical image synthesis
Synthetic data augmentation

ABSTRACT

Histopathological analysis is the present gold standard for precancerous lesion diagnosis. The goal of automated histopathological classification from digital images requires supervised training, which requires a large number of expert annotations that can be expensive and time-consuming to collect. Meanwhile, accurate classification of image patches cropped from whole-slide images is essential for standard sliding window based histopathology slide classification methods. To mitigate these issues, we propose a carefully designed conditional GAN model, namely *HistoGAN*, for synthesizing realistic histopathology image patches conditioned on class labels. We also investigate a novel synthetic augmentation framework that *selectively* adds new synthetic image patches generated by our proposed HistoGAN, rather than expanding directly the training set with synthetic images. By selecting synthetic images based on the confidence of their assigned labels and their feature similarity to real labeled images, our framework provides quality assurance to synthetic augmentation. Our models are evaluated on two datasets: a cervical histopathology image dataset with limited annotations, and another dataset of lymph node histopathology images with metastatic cancer. Here, we show that leveraging HistoGAN generated images with selective augmentation results in significant and consistent improvements of classification performance (**6.7**% and **2.8**% higher accuracy, respectively) for cervical histopathology and metastatic cancer datasets.

1. Introduction

Image analysis of digitized histopathological slides can contribute significantly to cancer diagnosis (Irshad et al., 2013). For instance, the diagnosis of cervical cancer and its precancerous stages can be accomplished through assessment of histopathology slides of cervical tissue by pathologists. An important outcome of the assessment is the cervical intraepithelial neoplasia (CIN) grade, an essential indicator for abnormality assessment identified by the abnormal growth of cells on the surface of the cervix. Over the past decade, computer-assisted diagnosis (CAD) algorithms have been developed for histopathology images to complement the opinion of the pathologist for accurate disease detection, diagnosis, and prognosis prediction (Gurcan et al., 2009). Considering the shortage of pathologists, automatic histopathology image classification systems have great potential in underdeveloped regions for its low cost and accessibility. Moreover, such a system can help pathologists with diagnosis and potentially mitigate the inter- and intra- pathologist variation.

The supervised training of image recognition systems often requires huge amounts of expert annotated data to reach a high level of accuracy. However, for many practical applications using histopathology images, only small datasets of labeled data are available due to annotation cost and privacy concerns, and the labels are often imbalanced between grades and subtypes. While traditional data augmentation can increase the amount of training data to some degree, commonly employed random transformations or distortions (such as cropping and flipping) lack flexibility and cannot fill the entire data distribution with missing data samples.

Motivated by the aforementioned difficulties in creating sufficiently large training sets for histopathology image recognition systems, we focus on the problem of expanding training sets with high-quality synthetic examples. Recently, several works in medical image analysis have leveraged unsupervised learning methods, more specifically, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), to mitigate the effects of small training

* Corresponding author.
E-mail address: sharon.x.huang@psu.edu (X. Huang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.media.2020.101816
1361-8415/

sets on network training (Liu et al., 2019; Frid-Adar et al., 2018). These works show that carefully designed GANs can generate visually appealing synthetic images, but two major issues remain insufficiently investigated for generalized and robust synthetic augmentation: 1) how to mitigate label ambiguity of generated images; and 2) how to ensure the feature quality of synthetic images used for data augmentation. In other words, blindly incorporating synthetic samples into the original training set, even if they are visually realistic, is not guaranteed to improve the classification model performance. Synthetic images without quality assurance can potentially adversely alter the data distribution and downgrade model performance. We provide a detailed analysis in Section 4.3.

In this paper, we aim at solving these two issues by designing a novel conditional GAN (cGAN) Mirza and Osindero (2014) framework, termed as HistoGAN, for high-fidelity histopathology image synthesis, then *selectively* adding synthetic samples generated by HistoGAN to the original training set. Our proposed HistoGAN model consists of multiple progressive generation and refinement modules which gradually generate images with better quality. To encourage the diversity of synthetic images, we incorporate the minibatch discrimination (Salimans et al., 2016) to reduce the closeness between examples inside a minibatch. Self attention (Vaswani et al., 2017) is employed to capture relationships between pixels inside an image. Such relationships contain crucial information about histopathology images, including the density distribution of nuclei and color changes in different locations. Class conditional batch normalization (De Vries et al., 2017) and spectral normalization (Miyato et al., 2018) are also utilized to stabilize the adversarial training process and improve the quality of synthetic images. Further, during HistoGAN training, we calculate a smoothed version of FrȨchet Inception Distance (FID) (Heusel et al., 2017) score after each epoch of training so that the trained models can be compared and the model with weights that give rise to the best FID score can be selected. Our proposed HistoGAN consistently generates realistic histopathology image patches on two different datasets, which shows the robustness and generality of the model.

Our proposed selective synthetic augmentation framework consists of two steps. First, we select generated images that can be classified into some class with certainty, by calculating the expectation of predictive entropy of each sample and keeping those samples with relatively low entropy (*i.e.*, high label confidence). Second, we compare the features of real images and synthetic images where the ground truth label of the real images matches the conditional label used to generate the synthetic images, and only select those synthetic images that are sufficiently close to the real-image centroid in feature space. The features of the images are extracted by a feature extractor pre-trained with Monte Carlo dropout (MC-dropout) (Gal and Ghahramani, 2016). This second step of selection is to ensure that a selected synthetic image indeed belongs to the class that corresponds to the conditional label used to generate it. The total number of selected samples is determined according to the augmentation ratio r (*i.e.*, the proportion of the number of augmented samples to the number of original training samples). Experimental results show that our proposed HistoGAN model along with selective synthetic augmentation significantly outperforms the baseline ResNet34 (He et al., 2016) model with traditional augmentation, and also outperforms the synthetic augmentation methods without selection.

To validate the effectiveness and generality of our proposed selective synthetic augmentation framework, we conduct extensive experiments on two histopathology datasets. We first study the 4-class (Normal, CIN 1-3) cervical histopathology image classification problem and evaluate our models on a heterogeneous epithelium image dataset (Xue et al., 2019) with limited and highly unbalanced numbers of patch-level annotations per class label. The second dataset we use is a small subset of the PCam dataset (Veeling et al., 2018), consisting of lymph node histopathology images. We compare our proposed selective synthetic augmentation method with baseline methods including baseline classification models, models trained with traditional augmentation, and models trained with synthetic augmentation but without quality-assuring selection. Experimental results show that our model achieves significant improvements with **6.7**% and **2.8**% higher accuracy than baseline classification models on cervical and lymph node datasets, respectively.

The main contributions of this work are as follows:

- We design a novel conditional GAN model architecture for synthesizing realistic histopathology image patches. A smoothed version of FID score is used as a metric to select the best cGAN model during training. With only a limited amount of training data, our GAN model can generate synthetic images with high fidelity and diversity.
- We propose a selective synthetic augmentation method that actively selects synthetic samples with high confidence of matching to their conditional label and are close to real images in feature space. By only adding selected synthetic samples instead of arbitrary synthetic samples to augment the limited training set, our proposed method can significantly outperform other baseline augmentation methods in improving classification performance. The proposed selective synthetic augmentation is general and can also be used in conjunction with other augmentation methods.
- We conduct extensive experiments on both a cervical histopathology dataset and a lymph node histopathology dataset. Compared with baseline models, including our previous state-of-the-art synthetic augmentation model (Xue et al., 2019), our proposed method improves the augmented classification performance.

2. Related work

2.1. Histopathology image classification

Machine learning, especially deep learning methods have achieved promising results on general histopathology image classification. While whole slide images (WSI) are often with unusually high resolutions, commonly used methods (Hou et al., 2016; Xu et al., 2017; Tomita et al., 2019) alleviate this issue by applying patch-level image classification on cropped image patches or sliding windows rather than the original WSI. Individual classification results on cropped patches are aggregated to infer the final image-level label for the WSI. In such methods, accurate patch-level image classification is fundamental to reach the accuracy level of human pathologists.

In the area of cervical histopathology analysis, existing literature (Chankong et al., 2014; Guo et al., 2016) have studied various supervised learning methods for nuclei-based cervical cancer classification. Chankong et al. (2014) proposed automatic cervical cancer cell segmentation and classification using fuzzy C-means (FCM) clustering and various types of classifiers. Guo et al. (2016) designed hand-crafted nuclei-based features for fusion-based classification on digitized epithelium histopathology slides with linear discriminant analysis (LDA) and support vector machines (SVM) classifier. While accomplishments have been achieved with fully-supervised learning methods, the training of models require large amounts of expert annotations of cervical histopathology images. Since the annotation process can be expensive, tedious, and time-consuming, it often results in limited or insufficient number of labeled data available for supervised learning models.

2.2. Conditional image synthesis

Generative adversarial networks (GANs) (Goodfellow et al., 2014) as an unsupervised learning technique, has enabled a wide variety of applications including image synthesis, object detection (Li et al., 2017) and image segmentation (Xue et al., 2018). Among variants of GANs, conditional GAN (cGAN) generates (Mirza and Osindero, 2014; Odena et al., 2017) more interpretable results with conditional inputs. For instance, images can be generated conditioning on class labels, which enables cGAN to serve as a tool to generate labeled samples for synthetic augmentation. Current state-of-the-art cGAN models often breaks the task into smaller gradual generation or refinement sub-tasks (Zhang et al., 2018; Karras et al., 2017), or employs large scale training (Brock et al., 2018), which enable them to generate high fidelity images. In this work, we use our proposed HistoGAN, which is inspired by state-of-the-art cGANs (Zhang et al., 2018; 2019; Brock et al., 2018), to generate high-fidelity synthetic images to augment classification model training. To improve the quality of synthetic images and stabilize the training process, our model utilizes numerous techniques including minibatch discrimination (Salimans et al., 2016), self attention (Vaswani et al., 2017), class conditional batch normalization (De Vries et al., 2017), and spectral normalization (Miyato et al., 2018) following prior art. While generating visually appealing histopathology images, HistoGAN serves as an essential prerequisite for the synthetic data augmentation.

2.3. Synthetic data augmentation

To better utilize training data and reduce over-fitting during the training process, data augmentation has become a common practice for training deep neural networks. The objective of augmentation is to add to the original training set new samples that follow the original data distribution. Therefore, a good augmentation scheme should generate samples that follow the original data distribution but are different from those in the original training set. On the other hand, a bad augmentation scheme can generate samples that deviate from the original data distribution thus can mislead training when added to the training set.

Traditional data augmentation (Wang and Perez, 2017) often involves transformations applied directly on original training data, such as cropping, flipping and color jittering. While serving as an implicit regularization, straightforward data augmentation techniques are limited in augmentation diversity. To overcome the limitation of traditional augmentation, several works have been done to improve the effectiveness of data augmentation. Rather than using a pre-defined augmentation policy, Auto Augmentations (Cubuk et al., 2019; Ho et al., 2019) use hyper-parameter searching to automatically find the optimal augmentation policy.

Another popular trend is to generate synthetic images to increase the amount and diversity of original training data, which we denote as *Synthetic Augmentation*. Along this direction, for natural images, Ratner et al. (2017) learns data transformation with unlabeled data using GANs. GAGAN (Antoniou et al., 2017) and BAGAN (Mariani et al., 2018) uses cGANs (Mirza and Osindero, 2014) generated samples to augment the standard classifier in the low-data regime. Compared with works done in the natural image domain, issues related to insufficient and imbalanced data are more prominent in the medical image domain. To mitigate these problems, researchers have been working on synthetic augmentation for medical image recognition tasks. Frid-Adar et al. (2018) proposes to use cGAN generated synthetic CT images to improve the performance of CNN in liver lesion classification. Gupta et al. (2019) synthesizes lesion images from non-lesion ones using CycleGAN (Zhu et al., 2017). Bowles et al. (2018) uses GAN derived synthetic images to augment medical image segmentation models. Zhao et al. (2018) proposes a GAN model for synthesizing retinal images from small sized samples and uses the synthetic images to improve semantic segmentation performance. Mahapatra et al. (2018) applies a Bayesian neural network (BNN) (MacKay, 1992) to calculate the informativeness of the synthetic images for improved classification and segmentation results. Zhao et al. (2019) uses transformations of labeled images for one-shot image segmentation. GAN based synthetic augmentation has achieved promising results, but typically blindly adds synthetic samples to the original data. Few consider how to assure the quality of synthetic images or control the augmentation step after image synthesis.

2.4. Our previous work

In our recent work (Xue et al., 2019), we propose a feature based filtering mechanism for synthetic augmentation. While improving classification performance, our previous cGAN generated images are not realistic enough and the work lacks rigorous study of its GAN model training and feature extractor training processes. In this work, we propose an improved GAN model for histopathology image generation, and develop a more general synthetic augmentation framework by reducing the randomness in GAN model and feature extractor training through MC-sampling and FID score based model selection. Our new contributions and differences from previous work are summarized as follows:

- We design and utilize an improved conditional GAN model architecture, namely HistoGAN, with a self-attention module among other techniques to stabilize the training and improve the quality of synthetic images.
- We propose a more general selective synthetic augmentation method which achieves better performances than our previous method.
- We conduct more comprehensive experiments including more ablation study and new results on the PCam dataset.

3. Methodology

In traditional fully-supervised training methods, the model is trained on training images and the inference is done by feeding the test data to the trained model. In previous GAN-based augmentation works (Frid-Adar et al., 2018; Madani et al., 2018), a GAN model is first trained to generate some synthetic images based on the training data, then the generated images are added to the original training data as a data augmentation strategy. However, since the discriminator in GAN only outputs a high level judgement (0 or 1) of the fidelity of generated images, such pipelines cannot guarantee that the generated data contain meaningful features which contribute to improving classification model training. To tackle this issue, we propose a selective synthetic augmentation algorithm to evaluate the quality and fidelity of synthetic images and select only those samples with high-confidence in label correctness and real-image likeness to be added to the training set. The comparison between different training procedures is illustrated in Fig. 1.

An overall illustration of our proposed selective synthetic augmentation method can be found in Fig. 2. We first train a conditional GAN model based on the labeled training images. The optimal model weights is selected based on the smoothed FID score Heusel et al. (2017). A pool of synthetic images are then generated using the selected model. All images are then passed into the image selection module to filter out the ones that fail to contribute sufficient amount of meaningful information. After image selection, a classification model is trained with both original and synthetic training data. Trained classification models can then be used for inference on test data. More details are introduced in the following subsections.

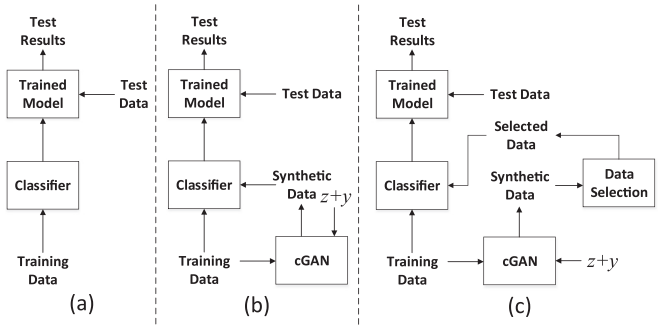

Fig. 1. Comparison between different training processes. (a) Traditional training pipeline; (b) Conditional GAN augmented training pipeline; (c) Our proposed selective synthetic augmentation with quality assurance. The input to the cGAN are noise vector z and label condition vector y.

3.1. HistoGAN model

In this section, we introduce our proposed HistoGAN architecture and how to select the best model with highest synthetic image quality from a set of trained models.

3.1.1. Model architecture

The conventional cGANs (Mirza and Osindero, 2014) have an objective function defined as:

$$\min_{\theta_G}\max_{\theta_D}\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x,y)] + \mathbb{E}_{z\sim\mathcal{N}}[\log(1-D(G(z,y)))]. \quad (1)$$

In the equation above, x represents the real data from an unknown image distribution P_{data} and y is the conditional label (*e.g.*, CIN grades). z is a random vector for the generator G, drawn from a standard normal distribution $\mathcal{N}(0,1)$. During the training, G and D are alternatively optimized to compete with each other.

Since there is no existing cGAN framework specifically designed for histopathology image synthesis, we choose to design a new model, *HistoGAN*, based on previous state-of-the-art conditional GAN models and techniques (Zhang et al., 2018; Brock et al., 2018; Zhang et al., 2019). The architecture of our HistoGAN model is illustrated in Fig. 3. We aim to generate synthetic images in a coarse-to-fine fashion through multiple stages, where details of images are gradually refined to guarantee the fidelity. The training procedure of HistoGAN is similar to Eq. (1). The generator of the first stage takes a random noise vector and class label as input, and the generator of remaining stages will take the output of the previous stage as input instead of random noise. To increase diversity among the generated examples and mitigate the issue of mode collapse indicated by the high homogeneity of the synthetic image pool, we incorporate the minibatch discrimination module (Salimans et al., 2016) into our discriminator. Following state-of-the-art works in conditional image synthesis (Zhang et al., 2019; Brock et al., 2018), class conditional batch normalization is used in both generators and discriminators to enhance the learning effectiveness of the inter class feature discrepancy. And spectral normalization (Miyato et al., 2018) is utilized in discriminators of all stages to further improve model performance.

To better capture the distribution of nucleus density and color changes in histopathology images of different classes, we leverage self attention (Vaswani et al., 2017; Zhang et al., 2019) at early stages of generation and throughout all stages in the discrimination process. The application of self attention mechanism enables both generator and discriminator to better learn the dependencies between spatial regions by looking at the relationship between one pixel and all other positions in the same image. Similar to Zhang et al. (2019), the image features from the previous hidden layer x are first transformed into two feature spaces q, k as query and key in self attention (Vaswani et al., 2017) to calculate the attention map. Let $q(x) = W_q x$ and $k(x) = W_k x$, the attention map over the ith location when synthesizing the jth region is

$$\alpha_{j,i} = \frac{\exp\left(s_{ji}\right)}{\sum_{i=1}^{N}\exp\left(s_{ji}\right)}, \text{ where } s_{ji} = \boldsymbol{q}(\boldsymbol{x}_i)^T\boldsymbol{k}(\boldsymbol{x}_j) \ . \quad (2)$$

The output of the self attention of the jth region o_j is calculated by applying attention weight over the value v as

$$\boldsymbol{o}_j = \sum_{i=1}^{N}\alpha_{j,i}\boldsymbol{v}(\boldsymbol{x}_i), \text{ where } \boldsymbol{v}(\boldsymbol{x}_i) = \boldsymbol{W}_v\boldsymbol{x}_i \ . \quad (3)$$

In all transformation matrices W_q, W_k, and W_v, weight matrices are implemented as 1×1 convolutions. Compared with the StackGAN model implemented in our previous work (Xue et al., 2019),

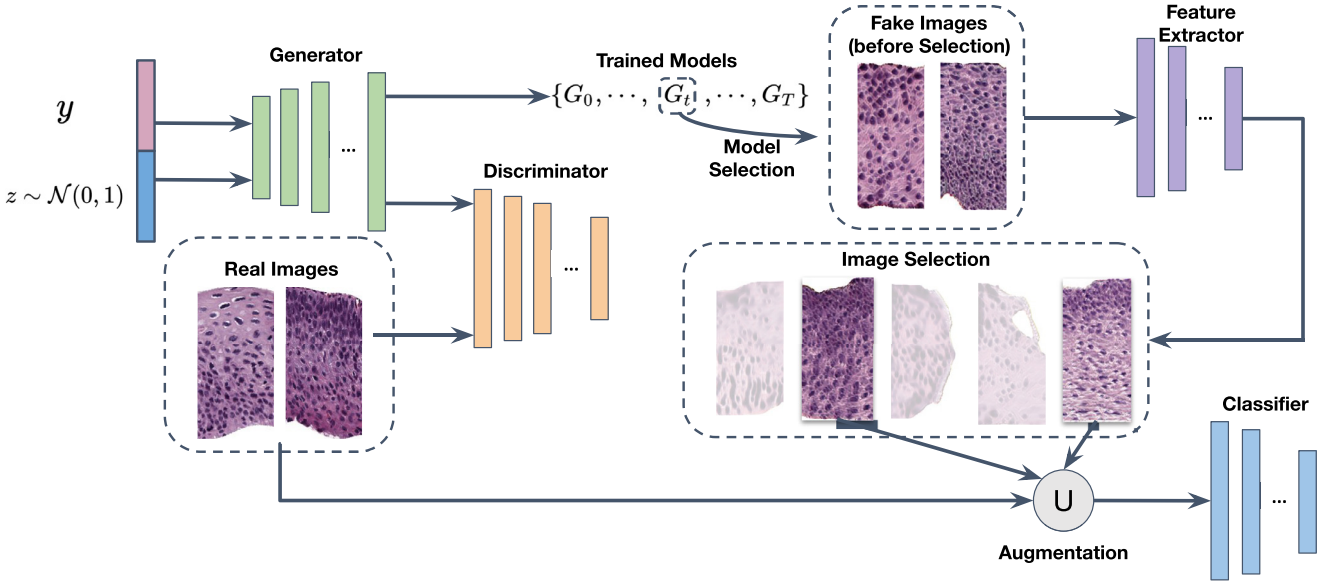

Fig. 2. The architecture of the proposed selective synthetic augmentation algorithm. The ∪ symbol indicates that the selected synthetic image set is unioned with the original training set to improve classification model training and test performance.

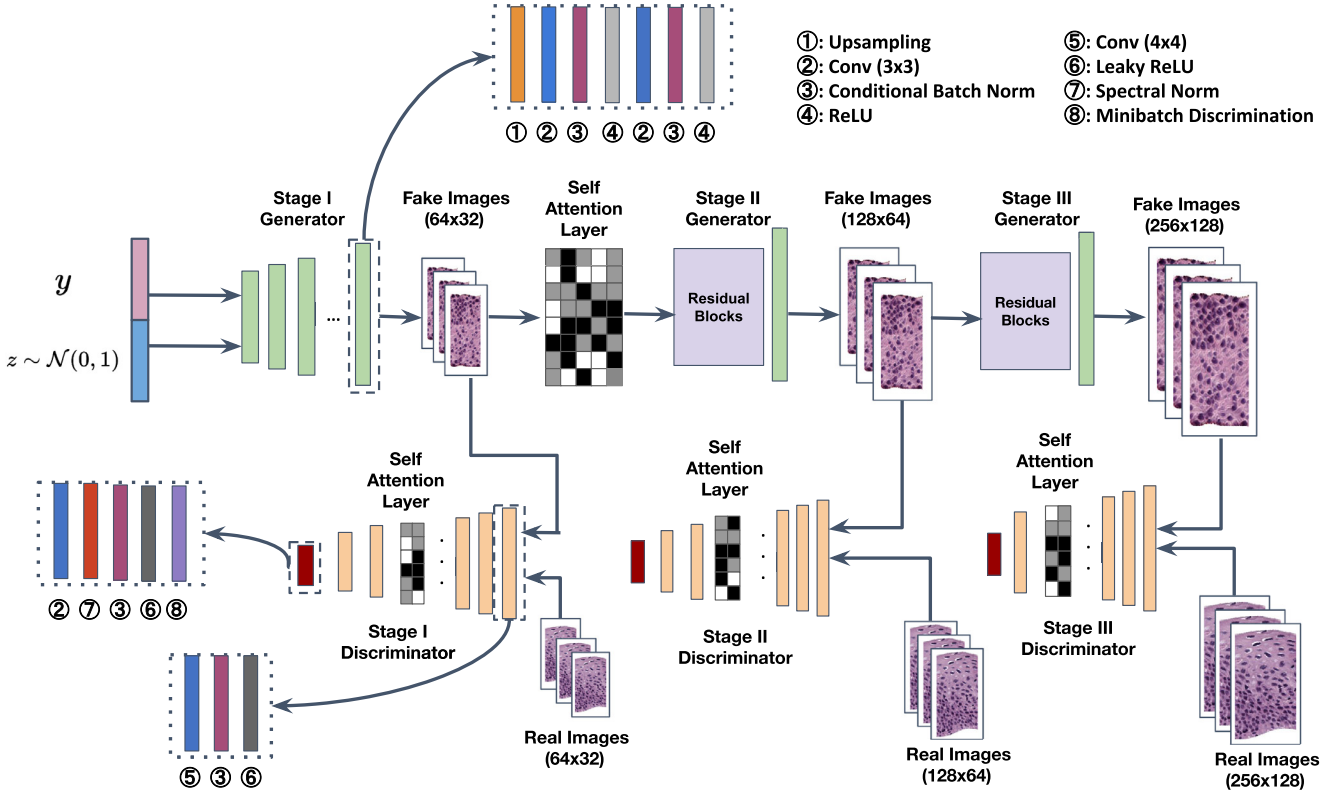

Fig. 3. The architecture of a 3-stage HistoGAN for cervical epithelium synthesis. The number of stages can be adjusted according to the desired final image resolution. Detailed features such as cytoplasm texture and nuclei shapes get progressively refined in synthetic images of higher resolution from stage I to III. The self attention layer is applied after stage I generator where the sketch outline and rough pattern of images are shaping up. Self attention layers are also incorporated in discriminators at all stages to further enforce the consistency of focused local regions more accurately. Conditional batch normalization De Vries et al. (2017) is used after convolutional layers for flexibly modulating convolutional feature maps.

our HistoGAN generates more realistic image patches which also benefits the following synthetic augmentation step. An example of cGAN result comparison is shown in Fig. 6.

3.1.2. Model selection

During training of the HistoGAN model, the model weights vary from epoch to epoch. A challenge is to determine which model weights gives rise to better synthetic image quality. For natural image synthesis tasks, Inception Score (Salimans et al., 2016) and Fŗechet Inception Distance (FID) (Heusel et al., 2017) score are two commonly used metrics. The calculation of these two metrics rely on the pre-trained Inception V3 (Szegedy et al., 2016) model trained on ImageNet (Deng et al., 2009). However, since the distribution of natural images and that of medical images such as cervical histopathology images can be quite different, we can not directly use the aforementioned two scores for evaluating our HistoGAN model. Instead, we follow the calculation of the original FID score while replacing the Inception V3 model pre-trained on ImageNet with a ResNet34 (He et al., 2016) model pre-trained on the cervical histopathology dataset.

To compare the trained models after running different numbers of epochs, we save the HistoGAN model after each epoch of training. To estimate the performance of each saved model, we calculate the FID score between the feature vectors of real and generated images extracted from the pre-trained ResNet34 model as follows:

$$d(x,\tilde{x}) = \left|\left|\mu_{x\sim P_{\text{data}}}\phi(x) - \mu_{\tilde{x}\sim P_G}\phi(\tilde{x})\right|\right|_2^2 + \mathbf{Tr}\Big(\Sigma_{x\sim P_{\text{data}}}\phi(x) + \Sigma_{\tilde{x}\sim P_G}\phi(\tilde{x}) - 2\big(\Sigma_{x\sim P_{\text{data}}}\phi(x)\Sigma_{\tilde{x}\sim P_G}\phi(\tilde{x})\big)^{\frac{1}{2}}\Big), \tag{4}$$

where $\tilde{x}$ represents synthetic images generated by the saved HistoGAN model being evaluated, and ϕ denotes the features extracted from intermediate layers of the pre-trained ResNet34 model. Assume feature vectors follow a multivariate Gaussian distribution, the mean and covariance are estimated for the real and fake data (Borji, 2019) for frȩchet distance calculation to measure the visual quality of generated images. Smaller FID scores indicate better visual quality. Although the FID score itself cannot guarantee agreement with human judgment, trends of FID often provide a reliable estimation of the quality of a GAN model. As we can observe from Fig. 4, due to the instability in GAN training, the FID scores of each saved epoch fluctuate constantly and fail to provide a distinguishable pattern. Based on the unaltered FID scores, one should choose the model saved at epoch 286 or epoch 374. However, one can see that images generated by these chosen models are not satisfactory as in Fig. 4. To get a robust estimation of model quality and mitigate the effect caused by outliers, we apply the Exponential Moving Average (EMA) (Hunter, 1986) algorithm to smooth the curve of original FID score. With smoothing, the FID score at time t is:

$$\hat{d} = \begin{cases} \mathrm{d}_t, & t = 1 \\ \alpha\hat{d}_{t-1} + (1-\alpha)\mathrm{d}_t, & t > 1 \end{cases} \tag{5}$$

We monitor the training process with the smoothed FID. As shown in Fig. 4, different values of α lead to different levels of smoothing in FID and we observed that the chosen model associated with the lowest smoothed FID score has better image quality than the model chosen using the lowest original FID score. In our experiments, we set α to 0.5 for a medium level of smoothing. One can see that, after smoothing with the EMA algorithm, the minimum in smoothed FID score is reached at epoch 634, which is the

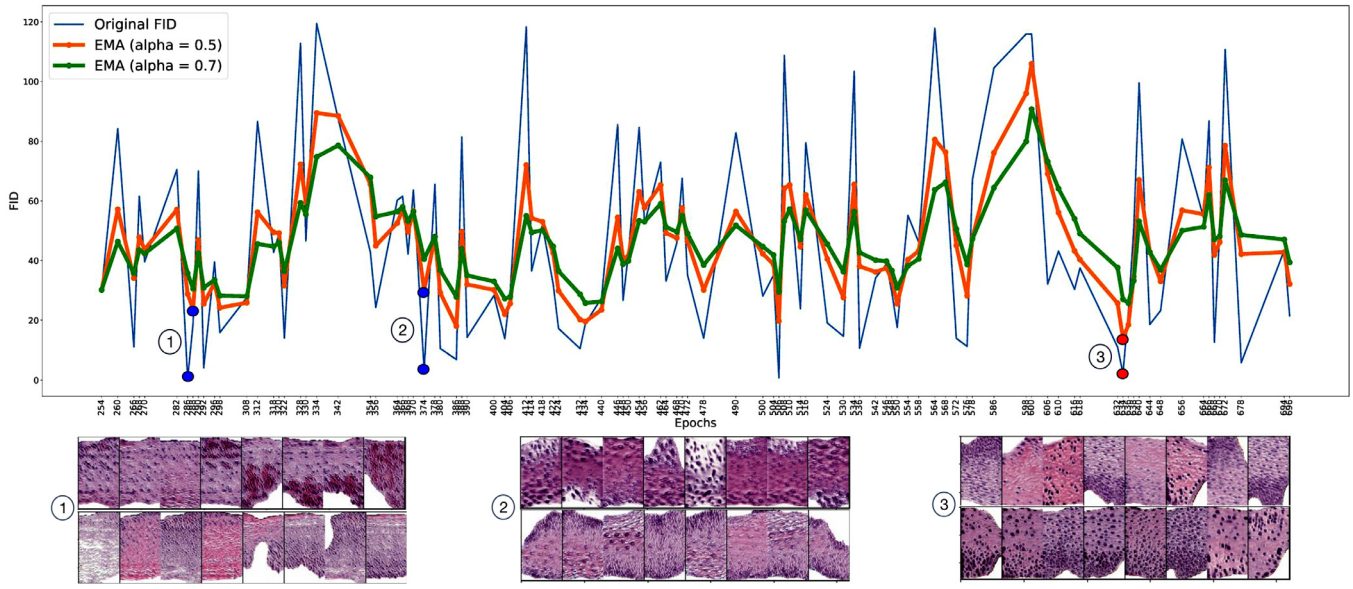

Fig. 4. FID scores (pre-trained ResNet34) of HistoGAN models saved after different number of epochs of training. Scores are smoothed with varying EMA parameter α.

GAN model we chose for the follow-on synthetic data augmentation.

3.2. Image selection

Given a trained cGAN model, one can sample infinite number of noise-vector inputs from the Gaussian distribution and generate infinite number of synthetic images. While a good cGAN model can generate images that look real, there are no guarantee that those images would be good to be used for augmenting the original training set in visual recognition tasks. In current GAN-based data augmentation methods, with different data augmentation ratio, different number of generated images are added to the training set. However, the effectiveness of such augmentation pipeline is heavily affected by the varying quality of synthetic images as well as the diversity of the images. To reduce the randomness in the synthetic augmentation process and selectively add in new images, we break the whole process into two steps: find samples that can be confidently classified into certain classes thus containing enough diagnosable features; then find samples whose features are within a certain neighborhood of class centroids in the feature space to assure matching between the synthetic image and its assigned label. Such steps are done with a pre-trained feature extractor to calculate centroids for real samples and extract features for fake samples. Considering that a single feature extractor cannot provide robust feature extraction results, we use a feature extractor with Monte Carlo dropout (MC-dropout) (Gal and Ghahramani, 2016) and take the expectation value of multiple samplings to reduce the uncertainty of feature extraction. A depiction of our proposed selective synthetic augmentation algorithm is shown in Fig. 5 and a detailed description is given in Algorithm 1.

The first step of selection is based on label certainty of a sample. In traditional machine learning systems, real samples that lie near the decision boundary are often assumed to contain more important features for classification purposes. However, as we conducted experiments to select good synthetic images, one interesting finding is that selecting the fake samples with more certain labels gives better classification performance than selecting those with less certain labels. This may be due to the cGAN model being imperfect and conditionally-generated fake examples with less label certainty being more likely to deviate from the real data distribution. In our algorithm, we evaluate the label certainty of a fake example by calculating the entropy score of its predicted class probabilities. If the feature extractor is certain that a sample can be classified into a certain class, the entropy score would be low. We rank the entropy scores of all generated images in ascending

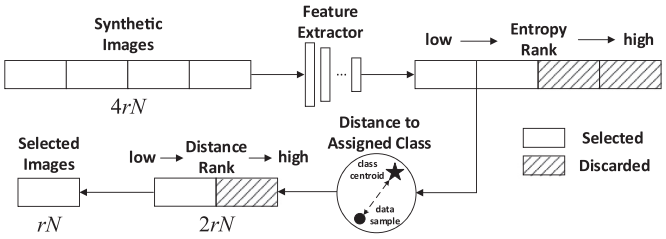

Fig. 5. Illustration of the image selection process. r and N represent the augmentation ratio and the number of original training data. The same feature extractor runs multiple times through MC-dropout for both entropy and class centroid distance calculations to increase robustness.

Algorithm 1 Selective Synthetic Augmentation

Input: a set of trained HistoGAN models $\{G_t\}$, number of classes $\mathcal{C}$, augmentation ratio r, number of original training samples $N = \sum_{i=1}^{\mathcal{C}} N_i$.
Output: selected synthetic samples $\mathcal{X}$ with $|\mathcal{X}| = rN$.
Initialization: $\mathcal{X}_1 = \emptyset$,$\hat{t} = \arg\min(\hat{d}_t)$,$G_{\hat{t}}$ generated samples $\mathcal{X}_0 = \{x_j^i : i \le \mathcal{C}, j \le 4rN_i\}$,entropy $\mathcal{E}^i = \{e_j^i : e_j^i = -\sum p_j^i \log p_j^i, i \le \mathcal{C}, j \le 4rN_i\}$.
for $x_j^i \in \mathcal{X}_0$ **do**
  if $e_j^i < \text{Median}(\mathcal{E}^i)$ **then**
    $\mathcal{X}_1 = \mathcal{X}_1 \cup \{x_j^i\}$
  end if
end for
class centroid distance $\mathcal{D}^i = \{d_j^i : d_j^i = D_f(x_j^i, c_i)\}$.
for $x_j^i \in \mathcal{X}_1$ **do**
  $d_j^i = D_f\{x_j^i, c_i\}$
  if $d_j^i < \text{Median}(\mathcal{D}^i)$ **then**
    $\mathcal{X} = \mathcal{X} \cup \{x_j^i\}$
  end if
end for

order and choose the first half of images with lower entropy. The necessity of this entropy-based selection is proved by experiments on different datasets, which will later be discussed in Section 4.

After the entropy selection step, we further select synthetic images based on their distance to class centroids in the feature space. In this second step of selection, all remaining samples that have passed the entropy-based selection will have their feature distances to their class centroids calculated. All distances will be sorted in ascending order and the first half of these samples with smaller distances will be kept. The motivation behind ranking samples based on their feature distance to class centroids is to help filter out samples whose assigned labels (i.e. the conditional labels used by the cGAN model to generate them) do not match their classified labels in feature space so that only samples that confidently match with their assigned labels are selected and added to the training set. In our implementation, instead of using a single run of the feature extractor to extract features, we run the feature extractor multiple times with MC-sampling and then calculate feature distances based on the average feature distance from the multiple runs. Similar to Xue et al. (2019), the feature distance between image x and centroid c is defined as

$$D_f(x, c_i) = \frac{1}{K}\sum_k \sum_l \frac{1}{H_l W_l}\left|\left|\hat{\phi}_l^k(x) - \hat{\phi}_l^k(c_i)\right|\right|_2^2 , \tag{6}$$

where $\hat{\phi}_l^k$ is the unit-normalized activation in the channel dimension A_l of the lth layer of the kth MC-sampling feature extraction network with shape $H_l \times W_l$. We denote the total sampling time as K. $D_f(x, c_i)$ can be regarded as an estimated cosine distance between sample and ith centroid in the feature space.

The centroid c is calculated as the average feature of all labeled training images in the same class. For class i, its centroid c_i is represented by

$$c_i = \left[\frac{1}{N_i}\sum_{j=1}^{N_i}\phi_1(x_j), ..., \frac{1}{N_i}\sum_{j=1}^{N_i}\phi_L(x_j)\right] , \tag{7}$$

where N_i denotes the number of training samples in ith class and x_j is the jth training sample. Similar to Eq. (6), ϕ_l is the activation extracted from the lth layer of the feature extraction network. L is the total number of layers utilized in the feature distance selection. c_i is retained by one time MC-sampling and fixed during the distance calculation.

In conclusion, given augmentation ratio r, we first generate $4rN_i$ images for each class i, then select rN_i images according to the two-step selection process described above. Regarding the choice of r, we provide an ablation study in Section 4.3.

4. Experiments

4.1. Datasets

The first dataset contains labeled cervical histopathology images collected from a collaborating health sciences center. All images are annotated by the same pathologist. The data processing follows (Xue et al., 2019), and results in patches with a unified size of 256 × 128 pixels. Compared with the dataset used in Xue et al. (2019), we include more data for more comprehensive experiments. In total, there are 1,284 Normal, 410 CIN1, 481 CIN2, 472 CIN3 patches. Examples of the images can be found in the first row of Fig. 6. We randomly split the dataset, by patients, into training, validation, and testing sets, with ratio 7:1:2 and keep the ratio of image classes almost the same among different sets. All evaluations and comparisons reported in this section are carried out on the test set.

To further prove the generality of our proposed method, we also conduct experiments on the public PatchCamelyon (PCam) benchmark (Veeling et al., 2018). PCam consists of 327,680 color patches extracted from histopathologic scans of lymph node sections with unified size of 96 × 96 pixels. The PCam dataset is split into 75%:12.5%:12.5% of training, validation, and testing sets, selected using a hard-negative mining regime. Each image is annotated with a binary label indicating presence of metastatic tissue. To mimic the situation where only a limited amount of training data is available, we use randomly selected 10% of the training set, which has 32,768 patches, to train our proposed HistoGAN model and the baseline classifier. Trained models are evaluated on the full test set.

4.2. Implementation details

4.2.1. HistoGAN implementation

The proposed HistoGAN model is trained in parallel on 4 NVIDIA TITAN Xp GPUs, each with 11G of RAM. We train HistoGAN with WGAN-GP Gulrajani et al. (2017) loss on the discriminators at all stages. Based on different sizes of images in the training set, we construct a 3-stage HistoGAN for cervical histopathology images and a 2-stage HistoGAN for the PCam lymph node histopathology images.

The input of the generator at the first stage is the concatenation of random noise and class label (*e.g.*, CIN1-3, Normal) that are first one-hot encoded and then embedded by a transposed convolution layer. The first stage generator consists of 4 up-sampling blocks with 3 × 3 conv kernels. Each block contains an upsample layer with bilinear interpolation followed by a combination of a convolutional layer with 3 × 3 kernel size. The output then goes through a conditional batch normalization De Vries et al. (2017) layer to modulate convolutional feature maps based on the corresponding assigned labels of the images generated. Blocks of the same architecture but different in and out channels are employed in generators of the next stages respectively, after a set of residual blocks.

Considering the future stages are learning the features from a more granularized level based on the output of the first stage, we employ self attention right after the first stage to facilitate the learning and focus on the desired features that are decisive for classification. Next, together with the real images from the original dataset with the same resolution, synthetic images of each scale are fed into corresponding stages of discriminators. Inside each discriminator, the main structure contains several down-sampling layers with 4 × 4 conv kernels. Similar to the aforementioned blocks in the generator, class-conditional batch normalization are used after each convolutional layer to embed more class specific information. The down-sampling layers are followed by a 3 × 3 conv layer, a spectral normalization layer, a batch normalization layer, a Leaky ReLU activation layer, a minibatch discrimination (Salimans et al., 2016) block for preventing mode collapse during GAN training, and a fully connected layer for the final output.

Regarding the hyperparameters, the HistoGAN model used for generating cervical histopathology and PCam images are trained with batch size set to 64 for the cervical and 256 for the PCam dataset for 1000 training epochs with fixed learning rate $2e-4$. The parameter δ for WGAN-GP loss is set to 50.

4.2.2. Model and image selection framework

In the next step, GAN models at each epoch are saved after the 100th epoch for model selection. For reasons mentioned in Section 3.1.2, the feature extractor used for FID score calculation is the same as our baseline classifier (ResNet34), followed by EMA-based smoothing to accentuate the pattern of synthetic image quality trend during the GAN training process. The optimal GAN model weights selected for further stages of our purposed sample selection corresponds to the epoch with the lowest adjusted FID score. Next, we generate $4rN_i$ synthetic images for each class i with

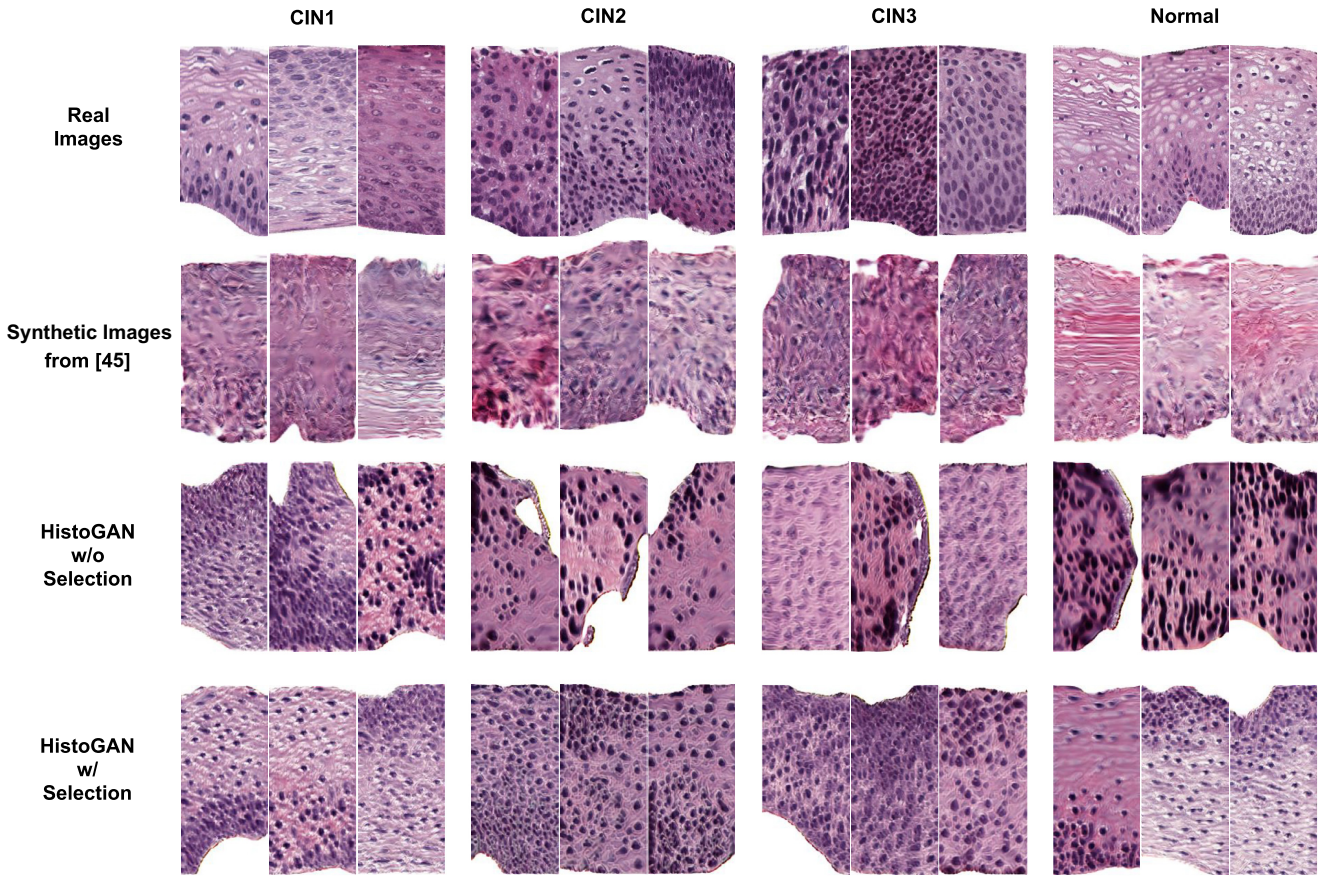

Fig. 6. Examples of real images, synthetic images generated from Xue et al. (2019), and images generated by our HistoGAN model trained on cervical histopathology dataset before and after selection. Our HistoGAN generates realistic images with clearly better visual quality than those by Xue et al. (2019). Zoom in for better view.

the chosen GAN, on which the same feature extractor is run for 5 times in order to extract the predicted probability from the softmax layer for entropy calculation, and also extract feature vectors after each residual block to obtain distance to centroids of ground truth. A dropout layer of rate 0.5 is inserted before the last residual block right above the fully-connected layer of the feature extractor (ResNet34) for Monte Carlo sampling. Then the generated images are ranked based on the mean of entropy across 5 runs in ascending order, of which half images in each class are kept. The selected pool of synthetic images are further ranked based on the mean of cosine distance to the centroid that corresponds to the assigned label of each image over 5 runs also in ascending order. Similarly, half are filtered out, leaving the rest for the final augmentation.

4.3. Results analysis

4.3.1. Evaluation by expert pathologists

To evaluate the quality of images generated by the proposed HistoGAN and validate the effectiveness of the selective synthetic augmentation method, we invited two pathologists to conduct expert evaluation on the cervical histopathology dataset. To prepare for the pathologist evaluation, we randomly chose 100 synthetic images where half of them are before selection and the other half are after selection. These images are then divided into 10 groups. Within each group of 10 images, there are two subgroups of 5 images where one subgroup is from the before-selection set and the other one is from the after-selection set. The 10 groups of images were then presented to the two pathologists who evaluated their quality independently. For each group, a pathologist was asked to choose one subgroup that has better quality, without knowing which subgroup corresponds to the one after selection; if the two subgroups were considered to have similar quality, the pathologist chose a tie. After the pathologists completed their evaluation, we compared their selected subgroups with the ground truth about which subgroups are from the after-selection image set. The comparison result shows that the two pathologists were able to differentiate before-selection subgroups from after-selection subgroups with high consistency: among the 10 groups, they chose the after-selection subgroup as having better quality 7 times, they chose a tie 2 times, and only once they chose the before-selection subgroup as having better quality. This evaluation result demonstrates that our image selection method is highly effective, since the expert pathologists consistently chose the after-selection images as having better quality.

Besides the group-level evaluation of our image selection method, the two pathologists also assessed the quality and realism of the individual synthetic images. They highlighted some realistic characteristics of the synthesized images, such as correct orientation, cell polarity, clear borders, and correct color of the cytoplasm. They also pointed out some unrealistic characteristics that repeatedly appeared in the generated image, such as smudged chromatin, missing nuclear details for large dark nuclei, and incorrect texture of large sheets of keratin. Despite the unrealistic aspects that they saw in the images, the pathologists actually view most of the images as containing meaningful features that make the images diagnosable. We are encouraged by these findings and plan to incorporate such expert knowledge in our future work to further improve our image synthesis model.

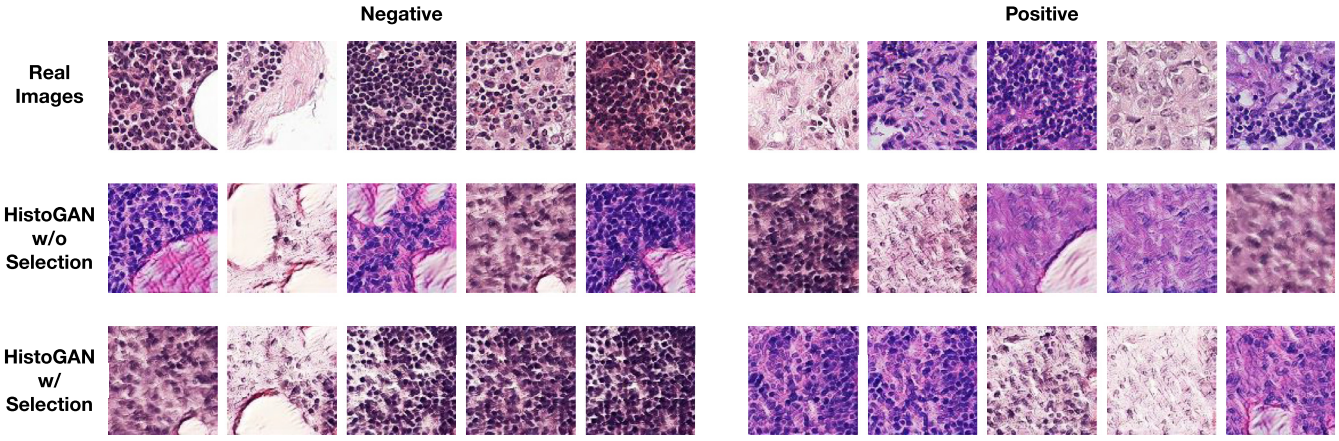

Fig. 7. Examples of real and synthetic images generated by HistoGAN trained on 10% of PCam dataset.

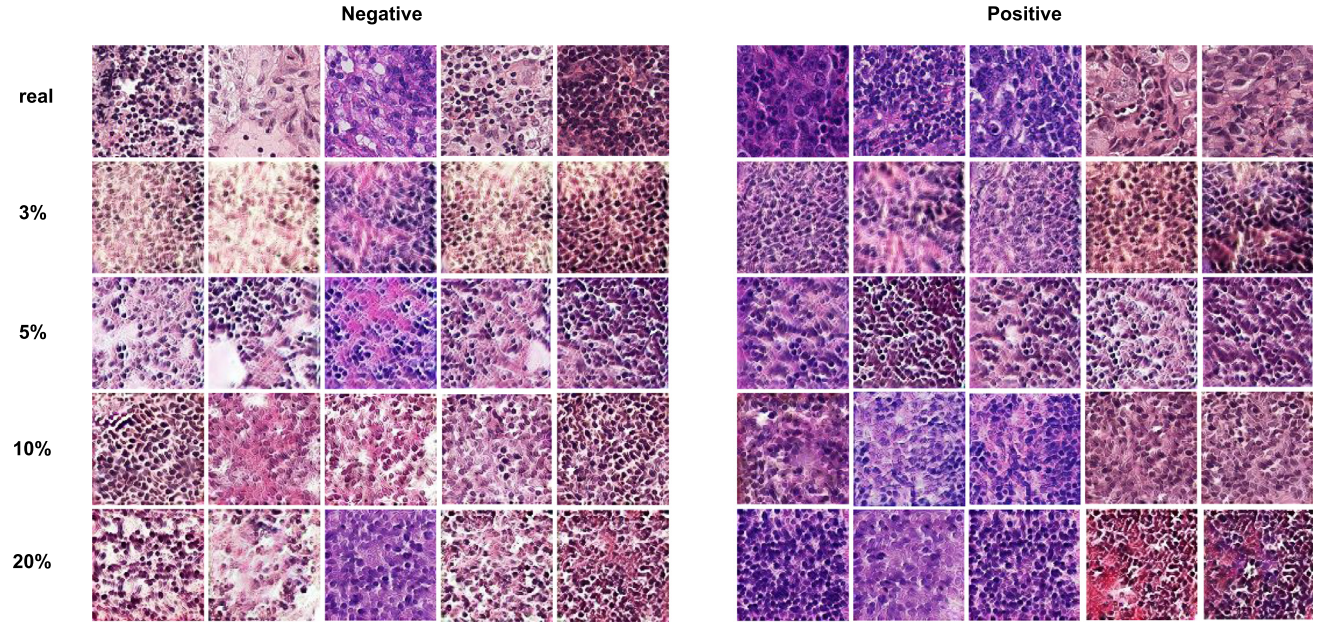

Fig. 8. Examples of real and synthetic images generated by HistoGAN trained on 3%, 5%, 10% and 20% of PCam dataset. All generated images are chosen from the pool after applying our proposed image selection method. Zoom in for better view.

4.3.2. Qualitative evaluation

The image synthesis results for cervical and lymph node datasets are demonstrated in Figs. 6 and 7, respectively. In Fig. 6, we also show a comparison of synthetic images generated by our previous work (Xue et al., 2019) and by our proposed HistoGAN in this work. In both datasets, as we have already achieved promising image generation results, determining whether those samples can be used for data augmentation or not cannot be easily done by human observations. However, the discrepancy between images with and without selection is much more prominent in the feature space. In order to visualize such differences, after training a baseline ResNet34 classifier with the original training data, we use the pre-trained ResNet34 model as the feature extractor to extract features from the last convolutional layer in the ResNet model. We explore the distribution of training samples, including both original images and synthetic images, in the feature space using t-SNE (Maaten and Hinton, 2008). In Fig. 10, without image selection, samples from different classes are entangled together, introducing obscuring noise that disrupts the data distribution that real data presents. On the contrary, selected images have clearly more distinguishable features and can potentially help with improving the classification model performance. Similar phenomenon is also observed with more noticeable pattern in Fig. 11: while data augmentation without image selection increases the number of training samples, the original data distribution is distorted. After image selection, the original data distribution is recovered along with more number of data points.

The self attention mechanism (Section 3.1.1) is a core improvement of our proposed HistoGAN model in this work as compared to the GAN model used in our previous work (Xue et al., 2019). In order to examine the role of self attention, we visualize the conditional attention maps for images from different classes in Fig. 9. From the figure, one can see that HistoGAN with self attention successfully learns meaningful features by attending to important areas containing patterns most useful in distinguishing images of different disease grades.

4.3.3. Quantitative comparisons

We report quantitative evaluation scores between all baseline augmentation models and our models including the accuracy, area under the ROC curve (AUC), sensitivity and specificity to provide a comprehensive comparison. All models are run for 5 rounds with

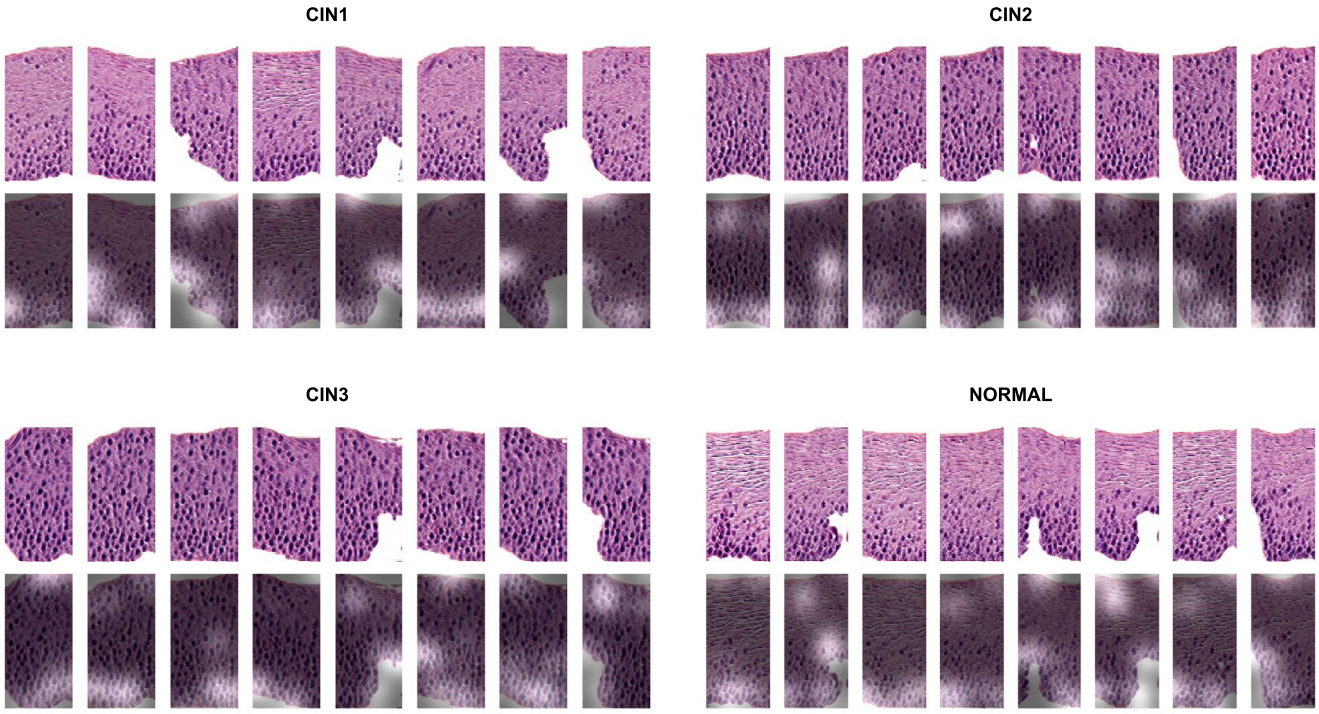

Fig. 9. The attention map extracted from the self attention layer applied after stage I generator as illustrated in Fig. 2. The first row shows the synthetic images generated by our proposed HistoGAN model; the second row gives the most attended regions of each image during the GAN training phase by overlaying the attention map on top of the original image. Higher attention scores correspond to the highlighted areas where distinguishing patterns like cell crowding and nuclei distribution are highlighted. It demonstrates the effectiveness of the attention mechanism incorporated in our HistoGAN model.

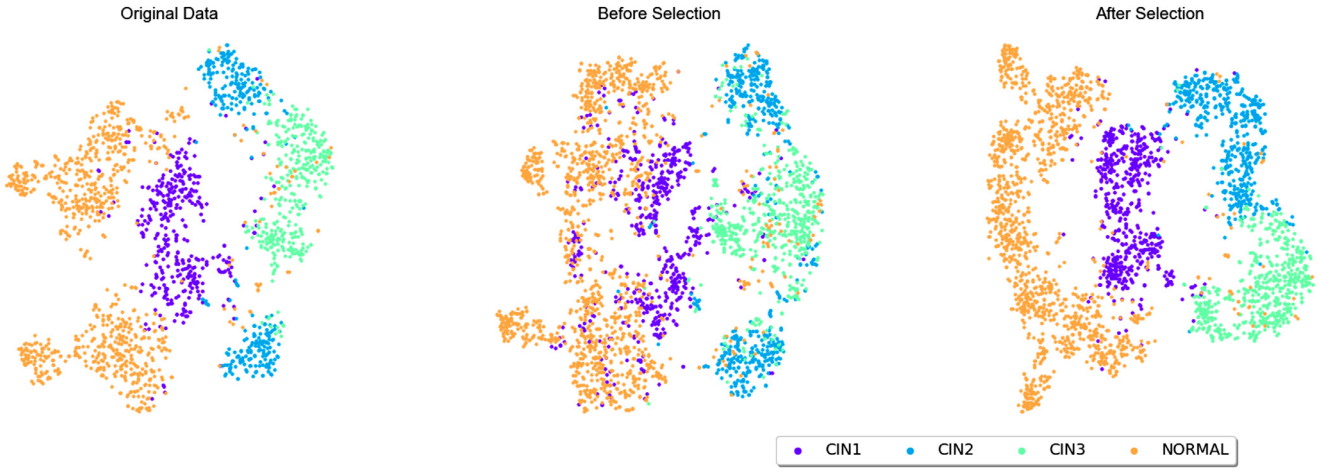

Fig. 10. t-SNE of the original and augmented cervical histopathology training set before and after image selection. The augmented training data after selection clearly have more distinguishable features than the ones without selection.

random initialization for fair comparison. The mean and standard deviation results of the 5 runs are reported.

In Table 1, we compare quantitative results with different baseline augmentation methods. We use the same backbone ResNet34 classifier with same hyperparameters setting in all experiments to ensure differences only come from the augmentation mechanisms. Beyond the backbone baseline model (He et al., 2016) without augmentation, we construct a baseline model with traditional augmentation including horizontal flipping and color jittering. Another baseline is GAN augmentation without selection where the training set is expanded by blindly adding GAN-generated images. We also compare the selective augmentation method proposed in this work with our prior work (Xue et al., 2019). Since in this work we use HistoGAN, an improved cGAN model that generates better synthetic images (as shown in Fig. 6) than the cGAN model originally described in Xue et al. (2019), we re-implemented (Xue et al., 2019) to also use HistoGAN generated images, for fair comparison of the image selection algorithms. From Table 1, one can see that the selective augmentation algorithm brings obvious benefits to all evaluation metrics, and our full model with augmentation ratio $r = 0.5$ achieves best performance in all metrics. More specifically, under $r = 0.5$, our image selection method improves the classification result by nearly 2% compared to the method in our prior work (Xue et al., 2019). This quantitative result demonstrates that our proposed selection method can better select high-quality images for augmentation than previous work.

To provide further insights on how the choice of the augmentation ratio r affects augmentation performance, we also conduct an

Fig. 11. t-SNE of the original and augmented PCam histopathology training set before and after image selection. While data augmentation without image selection increases the number of training samples, the original data distribution is distorted. After image selection, the original data distribution is recovered along with more number of data points.

Table 1
Classification results of baseline and augmentation models with different settings. Each model is run 5 times for the calculation of all evaluation metrics. For fair comparison between (Xue et al., 2019) and our work, we reimplemented (Xue et al., 2019) for it to use the same pool of synthetic images generated by HistoGAN.

	Accuracy	AUC	Sensitivity	Specificity
Baseline Model (He et al., 2016)	0.754 ± 0.012	0.836 ± 0.008	0.589 ± 0.017	0.892 ± 0.005
+ Traditional Augmentation	0.766 ± 0.013	0.844 ± 0.009	0.623 ± 0.029	0.891 ± 0.006
+ GAN Augmentation, $r = 0.5$	0.787 ± 0.005	0.858 ± 0.003	0.690 ± 0.014	0.909 ± 0.003
+ Single Filtering (Xue et al., 2019)*, $r = 0.5$	0.808 ± 0.005	0.872 ± 0.004	0.639 ± 0.015	0.912 ± 0.006
+ Selective Augmentation, $r = 0.5$	**0.821 ± 0.011**	**0.881 ± 0.007**	**0.671 ± 0.022**	**0.917 ± 0.005**

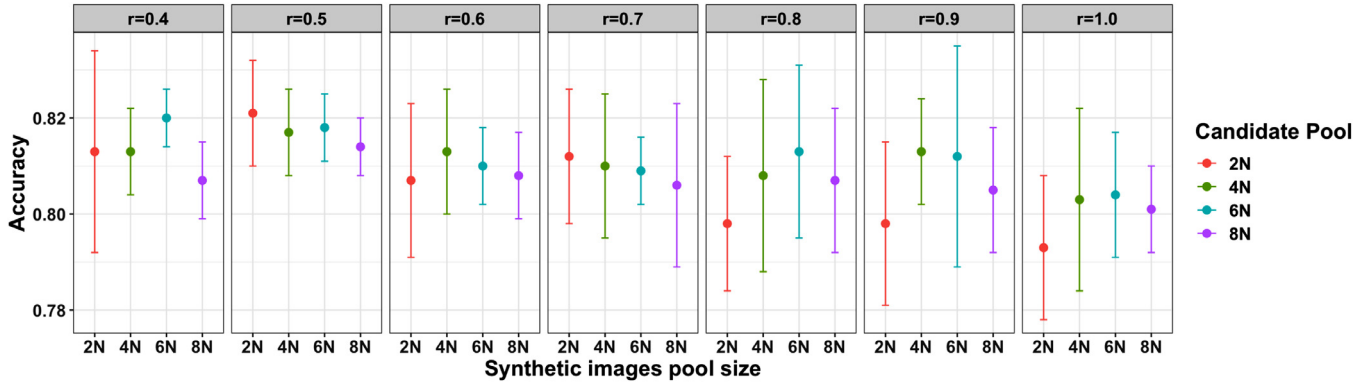

Fig. 12. Classification results of the proposed selective synthetic augmentation with different augmentation ratios on the cervical dataset. *N* in candidate pool sizes indicates the number of images in the original training dataset. For the same candidate pool size, selected images with different ratios are from the same pool. The error bar represents the standard deviation of classification accuracy from 5 multiple runs of each setting, the middle dot refers to the mean of 5 accuracy scores of the aforementioned multiple runs.

ablation study using different values of *r*, on different-sized candidate pools of HistoGAN-generated images. A summary of the ablation study is illustrated in Fig. 12. For this study, we generated synthetic image pools of four sizes: 2*N*, 4*N*, 6*N*, and 8*N*, where *N* is the size of the original training set. On these pools, we tested different values of *r*, between 0.4 and 1.0. Each test is run for 5 rounds, and the mean and standard deviation of the 5 runs are reported. From the results shown in Fig. 12, one can see that either too small or too large a value of *r* compromises the advantage of synthetic augmentation, and the best and most consistent performance gain is achieved at $r = 0.5$. This observation is true for all four pools of different sizes. Our explanation for this phenomenon is related to the motivation behind using selective synthetic augmentation: the synthetic images have different levels of quality, and the number of images with good quality and meaningful diverse features generated by a trained GAN model is limited. While our sample selection can provide quality assurance, the total number of diverse, good images that provide complementary information to the existing training set is constrained by the GAN model and more relevantly, by the original labeled training data used to train the GAN model. Therefore, a larger pool of generated images does not always translate to more high-quality images that will be selected by our method, as shown by this ablation study. Once our selection method has chosen those good images generated by the particular GAN model, adding more images such as images that do not improve diversity but may contain artifacts or bad features would indeed add noise to the training set thus degrade performance. Since our experiments show that the best aug-

Table 2
The performance of baseline and augmentation models using 3%, 5%, 10% and 20% of PCam as the training set. Each model is run 5 times for the calculation of all evaluation metrics. To further prove the effectiveness of our proposed selective augmentation, we compared the performance of our method and other augmentation methods when using 10% of PCam as the training set. 10% is chosen for demonstration because in this case the synthetic images show appealing visual quality as we can observe from Fig. 8, and consistently the classification performance presents improvement by a large margin.

PCam	Model	Accuracy	AUC	Sensitivity	Specificity
3 %	Baseline	0.872 ± 0.0030	0.914 ± 0.0019	0.826 ± 0.0080	0.903 ± 0.0030
	+ Selective Augmentation, $r = 0.5$	0.900 ± 0.0024	0.933 ± 0.0016	0.865 ± 0.0060	0.924 ± 0.0030
5 %	Baseline	0.893 ± 0.0006	0.929 ± 0.0004	0.863 ± 0.0010	0.913 ± 0.0020
	+ Selective Augmentation, $r = 0.5$	0.917 ± 0.0033	0.945 ± 0.0022	0.892 ± 0.0040	0.935 ± 0.0040
10 %	Baseline	0.910 ± 0.0012	0.940 ± 0.0009	0.883 ± 0.0050	0.929 ± 0.0030
	+ Traditional Augmentation	0.916 ± 0.0102	0.944 ± 0.0067	0.893 ± 0.0140	0.933 ± 0.0090
	+ GAN Augmentation, $r = 0.5$	0.920 ± 0.0020	0.947 ± 0.0014	0.898 ± 0.0050	0.935 ± 0.0020
	+ Selective Augmentation, $r = 0.5$	0.937 ± 0.0011	0.958 ± 0.0007	0.916 ± 0.0070	0.951 ± 0.0040
20 %	Baseline	0.932 ± 0.0014	0.955 ± 0.0010	0.909 ± 0.0080	0.948 ± 0.0040
	+ Selective Augmentation, $r = 0.5$	0.948 ± 0.0003	0.965 ± 0.0005	0.931 ± 0.0040	0.960 ± 0.0020

mentation performance is achieved at $r = 0.5$, we use this value for all ours and other baseline models and all experiments on the PCam dataset.

In Table 2, we use 3%, 5%, 10% and 20% of the training data in PCam to simulate training sets with limited annotations and evaluate our models on the full testing set. Compared with the cervical dataset, the baseline classification model achieves higher accuracy on the reduced PCam dataset which makes it more difficult to further improve the performance. However, our model still outperforms all baseline models using training sets of different sizes. For instance, when using 10% of the entire dataset as training data, the classification accuracy improved by 1% when using HistoGAN generated images for augmentation, without selection. After applying image selection, the accuracy is further improved by another 1.7%. By conducting experiments on two histopathology image datasets and showing improved classification performances, we prove that our proposed HistoGAN model and synthetic augmentation algorithm are general and can be applied to various types of histopathology data.

5. Discussion

Our proposed selective synthetic augmentation expands the training dataset by selectively adding synthetic images that do not distort the original data distribution, thus providing quality assurance in augmentation. The selected synthetic images are shown to improve the performance of automated image recognition systems with limited amount of manual annotation. We believe our proposed method is applicable to other histopathology image recognition tasks with insufficient annotated data. In addition, our proposed image selection algorithm is complementary to existing data augmentation methods, which further indicates the generality of our method.

While our selective synthetic augmentation significantly outperforms all baseline models, partial credits should go to the high-fidelity images generated by our proposed HistoGAN. However, the generated images are still not perfect, especially when viewed by expert pathologists, and we expect to further improve our GAN model with help from clinical experts. Besides the visual quality of images, the diversity of images also plays a critical role in synthetic augmentation. Since synthetic augmentation is imperative in scenarios with very scarce training samples, combining our pipeline with a GAN model that can learn from limited data (Wang et al., 2018; Lučić et al., 2019; Noguchi and Harada, 2019) would further improve the generality of our method. As we provide a solution to assure the synthetic image quality during augmentation, there is still room for improvement in selection mechanisms. More advanced methods for model selection and image selection, such as an end-to-end method and reinforcement learning based method, will be investigated in our future works.

6. Conclusion

In this paper, we design a new cGAN model termed HistoGAN for high-fidelity histopathology image synthesis and propose a synthetic augmentation method with quality assurance. By selectively adding realistic samples generated by HistoGAN into the original dataset, our method remarkably boosts the classification performance of baseline models. Experiments on two histopathology image datasets demonstrate the effectiveness and generality of our method.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yuan Xue: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Jiarong Ye:** Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Qianying Zhou:** Methodology, Software. **L. Rodney Long:** Resources, Data curation, Writing - review & editing. **Sameer Antani:** Resources, Data curation, Writing - review & editing. **Zhiyun Xue:** Resources, Data curation, Writing - review & editing. **Carl Cornwell:** Resources, Data curation, Writing - review & editing. **Richard Zaino:** Validation, Writing - review & editing. **Keith C. Cheng:** Validation, Writing - review & editing. **Xiaolei Huang:** Conceptualization, Funding acquisition, Project administration, Writing - review & editing.

Acknowledgments

This research is supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine, and Lister Hill National Center for Biomedical Communications. We gratefully acknowledge the help with expert annotations from Dr. Rosemary Zuna of the University of Oklahoma Health Sciences Center. We also thank Dr. Joe Stanley of Missouri University of Science and Technology for making the cervical histopathology data collection available.

References

Antoniou, A., Storkey, A., Edwards, H., 2017. Data augmentation generative adversarial networks. arXiv:1711.04340.

Borji, A., 2019. Pros and cons of GAN evaluation measures. Comput. Vis. Image Underst. 179, 41–65.
Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., Rueckert, D., 2018. GAN Augmentation: augmenting training data using generative adversarial networks. arXiv:1810.10863.
Brock, A., Donahue, J., Simonyan, K., 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv:1809.11096.
Chankong, T., Theera-Umpon, N., Auephanwiriyakul, S., 2014. Automatic cervical cell segmentation and classification in PAP smears. Comput. Methods Programs Biomed. 113 (2), 539–556.
Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2019. AutoAugment: learning augmentation strategies from data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 113–123.
De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C., 2017. Modulating early visual processing by language. In: Advances in Neural Information Processing Systems, pp. 6594–6604.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.
Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321, 321–331.
Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: NeurIPS, pp. 2672–2680.
Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, pp. 5767–5777.
Guo, P., Banerjee, K., Stanley, R.J., Long, R., Antani, S., Thoma, G., Zuna, R., Frazier, S.R., Moss, R.H., Stoecker, W.V., 2016. Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. IEEE J. Biomed. Health Inf. 20 (6), 1595–1607.
Gupta, A., Venkatesh, S., Chopra, S., Ledig, C., 2019. Generative image translation for data augmentation of bone lesion pathology. In: International Conference on Medical Imaging with Deep Learning, pp. 225–235.
Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147–171.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637.
Ho, D., Liang, E., Chen, X., Stoica, I., Abbeel, P., 2019. Population based augmentation: efficient learning of augmentation policy schedules. In: International Conference on Machine Learning, pp. 2731–2741.
Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433.
Hunter, J.S., 1986. The exponentially weighted moving average. J. Qual. Technol. 18 (4), 203–210.
Irshad, H., Veillard, A., Roux, L., Racoceanu, D., 2013. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review–current status and future potential. IEEE Rev. Biomed. Eng. 7, 97–114.
Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196.
Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S., 2017. Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1222–1230.
Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., Wang, Z., 2019. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. Engineering 5 (1), 156–163.
Lučić, M., Tschannen, M., Ritter, M., Zhai, X., Bachem, O., Gelly, S., 2019. High-fidelity image generation with fewer labels. In: International Conference on Machine Learning, pp. 4183–4192.
Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.
MacKay, D.J., 1992. A practical bayesian framework for backpropagation networks. Neural Comput. 4 (3), 448–472.
Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T., 2018. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: Medical Imaging 2018: Image Processing, vol. 10574. International Society for Optics and Photonics, p. 105741M.
Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M., 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 580–588.
Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C., 2018. BAGAN: Data augmentation with balancing GAN. arXiv:1803.09655.
Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv:1411.1784.
Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. arXiv:1802.05957.
Noguchi, A., Harada, T., 2019. Image generation from small datasets via batch statistics adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2750–2758.
Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 2642–2651.
Ratner, A.J., Ehrenberg, H., Hussain, Z., Dunnmon, J., Ré, C., 2017. Learning to compose domain-specific transformations for data augmentation. In: Advances in Neural Information Processing Systems, pp. 3236–3246.
Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.
Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., Hassanpour, S., 2019. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw. Open 2 (11), e1914645.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.
Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M., 2018. Rotation equivariant CNNs for digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 210–218.
Wang, J., Perez, L., 2017. The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Netw. Vis. Recognit. 11.
Wang, Y., Wu, C., Herranz, L., van de Weijer, J., Gonzalez-Garcia, A., Raducanu, B., 2018. Transferring GANs: generating images from limited data. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 218–234.
Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinform. 18 (1), 281.
Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SegAN: adversarial network with multi-scale l 1 loss for medical image segmentation. Neuroinformatics 16 (3-4), 383–392.
Xue, Y., Zhou, Q., Ye, J., Long, L.R., Antani, S., Cornwell, C., Xue, Z., Huang, X., 2019. Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 387–396.
Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019. Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363.
Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2018. StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach.Intell. 41 (8), 1947–1962.
Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8543–8553.
Zhao, H., Li, H., Maurer-Stroh, S., Cheng, L., 2018. Synthesizing retinal and neuronal images with generative adversarial nets. Med. Image Anal. 49, 14–26.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.