Self-Supervised Representation Learning using Visual Field Expansion on Digital Pathology

Joseph Boyd[1], Mykola Liashuha[1], Eric Deutsch[2], Nikos Paragios[3], Stergios Christodoulidis[*1], and Maria Vakalopoulou[*1]

[1]MICS Laboratory, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
firstname.lastename@centralesupelec.fr
[2]Department of Radiotherapy, Gustave Roussy Cancer Campus, 94800 Villejuif, France
eric.deutsch@gustaveroussy.fr
[3]Therapanacea, 75014 Paris, France
n.paragios@therapanacea.eu

Abstract

The examination of histopathology images is considered to be the gold standard for the diagnosis and stratification of cancer patients. A key challenge in the analysis of such images is their size, which can run into the gigapixels and can require tedious screening by clinicians. With the recent advances in computational medicine, automatic tools have been proposed to assist clinicians in their everyday practice. Such tools typically process these large images by slicing them into tiles that can then be encoded and utilized for different clinical models. In this study, we propose a novel generative framework that can learn powerful representations for such tiles by learning to plausibly expand their visual field. In particular, we developed a progressively grown generative model with the objective of visual field expansion. Thus trained, our model learns to generate different tissue types with fine details, while simultaneously learning powerful representations that can be used for different clinical endpoints, all in a self-supervised way. To evaluate the performance of our model, we conducted classification experiments on CAMELYON17 and CRC benchmark datasets, comparing favorably to other self-supervised and pre-trained strategies that are commonly used in digital pathology. Our code is available at https://github.com/jcboyd/cdpath21-gan.

1. Introduction

The characterization and quantification of tissue using microscopy images is considered the gold standard for diagnosis and prognosis in evaluating treatment response for patients with cancer. Traditionally, clinical pathologists examine thin tissue slices under a microscope identifying known biomarkers such as cancer cells, cancer subtypes, percentage of tumour infiltrating lymphocytes, and others. These processes are tedious and time consuming, and moreover may suffer from inter- and intra-observer variability. Currently, with the recent efforts of the computational pathology community, the digitization of tissue slides to whole slide images (WSI) fit for automated analysis is rapidly growing, while more and more research is focused on developing algorithms that can provide accurate and robust tools for clinical use. In particular, with the recent advances in deep learning the automatic analysis of WSIs under supervised and weakly supervised schemes have become very popular [9, 5, 25, 36].

Although there has been considerable progress in recent years towards the automatic processing of WSIs and its use in clinical practice, there are still a number of lingering challenges. Firstly, the gigapixel size of WSIs makes the development of tailored machine learning techniques challenging. To address this issue, the processing is mainly performed on a tile level while multiple instance learning (MIL) schemes are often developed to predict different clinical endpoints [40, 25, 38]. Furthermore, the size of WSIs makes the annotation process by human experts difficult and time-consuming. Most current methods require some sort of supervision either in the form of fully supervisory sig-

*These authors contributed equally to this work.

nals or by utilizing some weakly supervised scheme, making annotated data essential for the development of robust algorithms. Furthermore, with the shift to deep learning the availability of annotations can also impact the generated representations, limiting the reported performances. On these grounds, some recent methods consider the use of pre-trained representations, usually obtained by ImageNet pre-training without investing in generating representations that are specific to WSI images [12].

Unsupervised or self-supervised approaches have recently been studied as an alternative to fully supervised methods, with promising results. Such methods can eliminate the need for annotations and as such can greatly increase the amount of effective training data. With an appropriate problem formulation, self-supervised or unsupervised signals can be leveraged so as to extract compact and informative representations [30, 33, 17]. Traditionally, however, these methods report lower performance than fully supervised ones making their use at present less popular, as they cannot be integrated into clinical practice.

In this study, we propose a novel self-supervised generative method for realistic expansions of the field of view of histopathology image tiles, outpainting the visual context with plausible structure and details (Figure 1). Our method utilizes a progressively grown model that learns robust representations using adversarial training, extrapolating the margins of a tile in a realistic way. In such a setup the network learns in a self-supervised manner the structures and objects that occur in different tissue types within WSIs. The contributions of this work are: *(i)* a novel self-supervised, generative scheme for extrapolating an extended visual field of the tile; *(ii)* a method to generate artificial tiles given specific structures; and *(iii)* generation of histopathology representations that outperform other state-of-the-art pre-trained and self-supervised strategies for histopathology classification tasks. To the best of our knowledge, this is the first attempt to explore the use of expansion of the visual field on histopathology slides. Our method outperforms other state-of-the-art and commonly used methods for feature representation in histopathology, reporting performances close to comparable fully supervised methods.

2. Related work

Expansion of image borders for semantic image extrapolation is a problem that has previously been investigated by the computer vision community. However, it has not widely been explored by the biomedical imaging community, and in particular within digital histopathology. More specifically, [39] proposed an encoder-decoder framework with skip connections and recurrent content transfer for generative natural image scenery prediction. Moreover, [37] proposed a semantic regeneration network based on deep generative models for wide-context semantic image extrapolation on natural images. Our method shares similar ideas with these works, however it is based on progressively grown adversarial autoencoders for the specific task of learning robust feature representations of histopathological images.

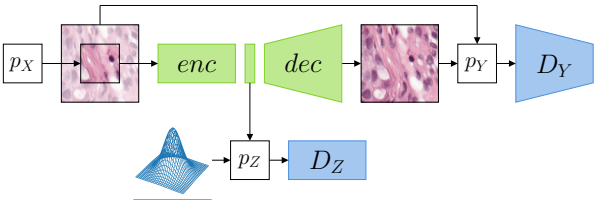


Figure 1: Learning representations through visual field expansion. An adversarial autoencoder with a progressively grown decoder [20] and two adversarial losses that ensure a Gaussian distribution for the latent code (D_Z) and a realistic visual field expansion (D_Y).

In recent years, generative models, and in particular generative adversarial networks (GANs), have been widely used in histopathology [35]. In [16] the authors proposed a GAN-based method for histopathology image segmentation that synthesizes heterogeneous sets of training image patches of different tissue types. The proposed method enumerates possible ground truth structures during generation of synthetic training patches. In the event the resulting patch is not realistic, the framework decreases its impact in the training loss while, if both realistic and also rarely synthesized, then its impact in the training loss is increased. The method has been validated for the task of nuclei segmentation, proving better generalization compared to other state of the art methods without training data. Recently [8] proposed the use of GANs to capture key tissue features, structuring these characteristics in its latent space. The authors based their framework on [19, 4, 21] and they show that their model induces an ordered latent space based on tissue characteristics (e.g. cancer cell density or tissue type), allowing to perform linear vector operations that correspond to high-level tissue tile changes. Contrary to these methods, our framework is based on an encoder-decoder architecture generating high-quality visual content from given structures. Finally, other tasks in which GANs are regularly utilized in histopathology include stain normalization and domain adaptation [3, 10], mainly using the CycleGAN architecture [41].

The use of self-supervised or pre-trained schemes in histopathology is a common strategy since the amount of annotations is usually limited. Different types of self-supervised problems have been proposed in the literature [32], focusing on super-resolution and color normalization [24], magnification prediction [30] as well as contrastive learning [7]. In particular, [23] proposed the use

of different self-supervised domain-specific auxiliary tasks (Self-Path) such as magnification, jigmad and hematoxylin channel prediction as well as domain-agnostic tasks such as adversarial losses, augmentations and domain prediction. Self-Path reports similar or better performance to supervised baselines on the CAMELYON16 dataset in the case of a low number of annotations. Our work explores another alternative to this direction, learning to predict the invisible while generating robust representations of histopathology images.

3. Method

Our model is trained according to the self-supervised task of visual field expansion. Formally, given a target image $\mathbf{x} \in X$ of dimension $w \times h \times C$, we aim to expand the image artificially into a larger image $\mathbf{y} \in Y$ of dimension $W \times H \times C$, where $W > w$ and $H > h$ and such that for some contiguous crop of $\mathbf{y}$, denoted $\mathcal{C}(\mathbf{y})$, we have it that $\mathcal{C}(\mathbf{y}) \approx \mathbf{x}$. In practice, we take the central crop having half the size of the target image in each spatial dimension (thus, $1/4$ of the target image area), and the model is trained to expand from the center outwards. We hypothesise that a model trained for visual field expansion necessitates a rich representation for the observed tissue properties of the input, thereby yielding a powerful encoder of histopathology tiles.

Our proposed model leverages the adversarial autoencoder framework [27]. An overview of our model is presented in Figure 1. The model consists of an encoder $enc(\cdot) : X \to Z$ mapping image crops $\mathbf{x} \in X$ to a latent vector representation $\mathbf{z} \in Z$, and a decoder $dec(\cdot) : Z \to Y$ mapping the latent representation to an expanded image $\mathbf{y}$. The decoder is furthermore a generative model, as the latent representation is trained adversarially against a discriminator network, $D_Z : Z \to \{0, 1\}$ to resemble a known template distribution p_Z . We write,

$$\min_{enc,dec} \max_{D_Z} \mathcal{L}_{AA} = \mathbb{E}_{\mathbf{z} \in p_Z}[\log D_Z(\mathbf{z})] + \mathbb{E}_{\mathbf{x} \in p_X}[1 - \log D_Z(enc(\mathbf{x}))] + \mathbb{E}_{\mathbf{y} \in p_Y}[\lambda \cdot R(\mathbf{y}, dec(enc(\mathcal{C}(\mathbf{y}))))] \quad (1)$$

where $R(\cdot, \cdot)$ is a reconstruction error function to ensure the autoencoding property and λ a manually tuned weight. In practice, we choose R to be the L_1 loss, given its aptitude for sharp image synthesis in image-to-image translation tasks [18, 41]. In addition to the convention of alternating generative and discriminative steps, the autoencoder and encoder objectives are trained in alternating reconstruction and regularisation steps. We include an additional discriminator $D_Y : Y \to \{0, 1\}$ on the decoder output images. This proved crucial to ensuring a consistency among the fine details in the original and extrapolated regions. Our full training objective therefore becomes,

$$\min_{enc,dec} \max_{D_Z, D_Y} \mathcal{L} = \mathcal{L}_{AA} + \mathbb{E}_{\mathbf{y} \in p_Y}[\log D_Y(\mathbf{y})] + \mathbb{E}_{\mathbf{x} \in p_X}[1 - \log D_Y(dec(enc(\mathbf{x})))] \quad (2)$$

To facilitate the generation of high-resolution images, we adopt the progressive growing algorithm for GANs [20]. This training algorithm divides model training into stages, beginning with a low resolution target and "fading in" targets of increasingly high resolution in each successive stage in a form of curriculum learning. Pioneer networks [14] are an example of a progressively grown generative autoencoder, however, their training is not GAN-based, and they grow both encoder and decoder symmetrically. In contrast, our model has a fixed encoder taking a fixed 112×112 image crop, which we decode progressively.

4. Experimental setup

4.1. Datasets

In order to evaluate and compare the developed models with the baselines we utilize two publicly available datasets, CAMELYON17 [2] and CRC[1] [22]. CAMELYON17 consists of 100 patients, each providing 5 WSIs. A WSI contains one or more sections extracted from axillary lymph nodes. Of the 500 slides only 50 contain metastasis, with the remaining 450 consisting of negative samples.

Our pipeline for the CAMELYON17 WSIs follows a standard preprocessing, similar to that described in [25]. In practice, the tissue area is first extracted from each WSI in an unsupervised way by converting the WSI to the HSV color space and by using an adaptive Otsu threshold on the saturation channel. Using the resulting masks we crop all tiles of size 224×224px in a regular grid within the tissue area at 40X magnification. From the complete set of tiles, all tiles within metastatic regions were selected and an equal number of normal tiles were randomly sampled. If a tile consisted of more that 50% background pixels it was discarded. We defined positive samples to be a tile that has at least one tumour pixel inside its central 86×86px region. A patient-wise splitting was then performed to split the tiles into training, validation and test sets ($70/15/15\%$ patient-wise). In total, the final training, validation and test sets were respectively 140k, 30k and 15k tiles with equal numbers of positive and negative samples.

The CRC dataset consists of a set of 100k non-overlapping image tiles from H&E stained WSIs of human colorectal cancer (CRC) and normal tissue. Train and test splits are provided. We randomly sampled without replacement a validation set amounting to 10% of the training data.

[1] https://doi.org/10.5281/zenodo.1214456

All images are 224×224px at 0.5 microns per pixel (MPP) and are color-normalized using Macenko's method [26]. The tiles are classified into nine tissue type classes: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

4.2. Implementation and training details

To ensure a fair comparison, the encoders of all our models were fixed to be a Resnet18 [13] with pretrained ImageNet weights. The sole architectural differences between the encoders were the encoder heads, that is, the additional layers appended to each encoder to fulfill the respective training objectives. For our proposed model, this consisted of an additional convolutional and fully-connected projection layer, immediately following the final convolutional layer of the Resnet18 backbone. Given that the generative model is designed to expand the spatial dimensions from input to output, a larger linear "deprojection" layer is used in the decoder. Note that the decoder is not used in the downstream feature extraction tasks.

By inspection, we found that freezing the early residual layers of the ResNet18 backbone improved the quality of the model outputs. We manually tuned the freezing strategy to incorporate the first three convolutional blocks, leaving only the final block for fine tuning at training time. Furthermore, to curb the deleterious effects of the large gradients of the randomly initialized model head at the start of training, we lock all backbone weights for the first epoch of training, after which the models are fine-tuned. We used the Adam optimizer ($lr = 0.001, \beta_0 = 0, \beta_1 = 0.999$) for the training of our proposed method. The progressive training proceeded over five stages of resolution doubling, from 7×7px to 224×224px. In the first half of each stage, the new resolution was faded in linearly as a weighted sum of the bilinearly upsampled former and new image resolutions, as in [20]. On average, each stage consisted of the equivalent of approximately 80 epochs of the training data, with mini-batches of size of 16.

A computational cluster with NVIDIA Tesla V100 GPUs was utilized for all experiments while PyTorch [28] and TensorFlow [1] were used for model implementation[2] and training. The total training time was approximately 36 hours.

5. Results

5.1. Evaluation of generated images

A number of qualitative examples of generated image expansions for CAMELYON17 tiles are presented in Figure 2. On the left we present the inputs to the model together with the ground truth expanded region, and on the right we present the outputs of our model. One may observe that our network is able to realistically generate specific global structures, in addition to fine details, closely approximating the hidden regions. Indeed, the structure and content of the tiles are presented and reproduced accurately. In supplementary materials we separately provide a video of tiles sampled directly from the decoder from template Gaussian noise, over the course of the progressive training.

[2]Code available at https://github.com/jcboyd/cdpath21-gan

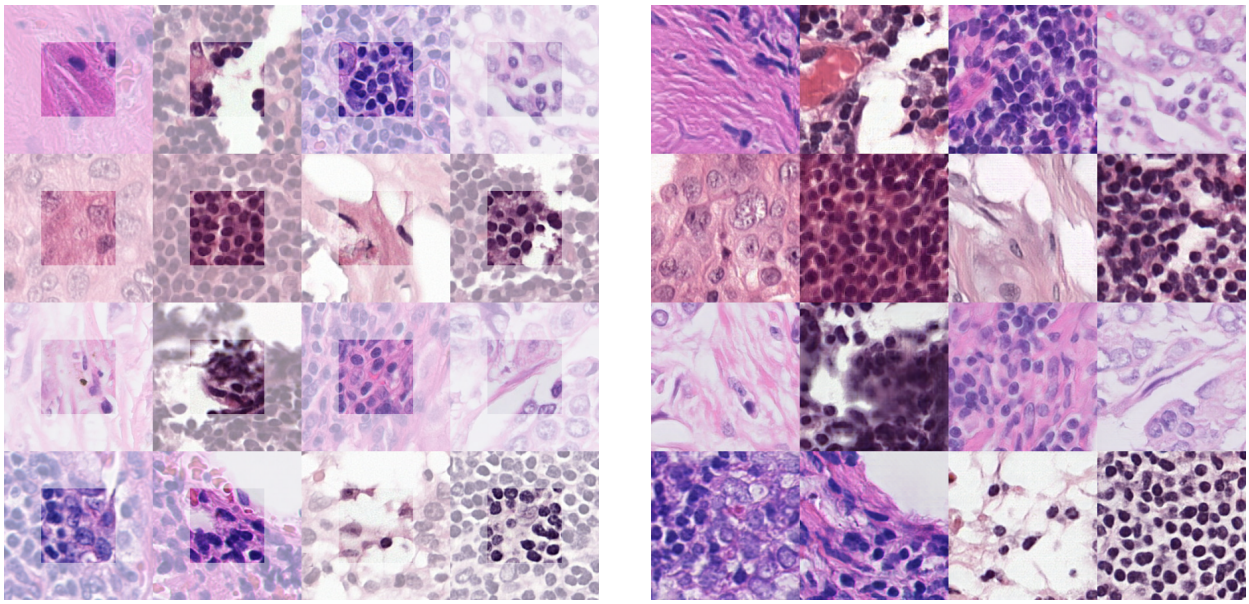
	CAMELYON17	CRC
FID sampled	33.37	50.47
FID expanded	20.76	37.05

Table 1: FIDs for generated and expanded tiles for the proposed model on the CAMELYON17 and CRC datasets.

Aside from the qualitative evaluation of the generated images, we calculate the Fréchet Inception Distance (FID). FID evaluates the quality of the image outputs of a generative model. In practice, FID compares the feature representations generated by an Inception network for real and generated images. We use the `pytorch-fid` library [31] to evaluate the quality of both generated images and image expansions. In the first case, we generate synthetic images by randomly sampling 512-dimensional Gaussian noise and decoding it directly into a histopathology tile with the decoder network. As recommended by [15] we compute FID between 50k generated images and all training images. In the second case, input crops are fed through the full model to create a set of image expansions. These are used in place of sampled images and the FID is computed as before. Table 1 shows the results of these two evaluations on each of the two datasets. In particular, we note a better performance for expansion than generation. This suggests some degree of divergence in the target and learned distributions of the latent variables. We hypothesise that a small amount of non-Gaussian information, undetected by D_Z, is encoded by enc to achieve better reconstructions. We thus characterise a tension between the training objectives of accurate reconstruction and the distributional properties of the latent space. Finally, we observe that our reported FIDs are lower in the case of CAMELYON17 (breast cancer) than the CRC (colorectal cancer) something that is in accordance with the experiments reported in [8] for the same cancer types.

5.2. Evaluation of learned representations

To assess the quality of the learned representations, we used the latent codes $\mathbf{z}$ of our models as features in a pair of downstream classification tasks. For the CAMELYON17 dataset, this was the binary classification of tiles into metastatic and non-metastatic classes; for CRC, this was the classification of tiles into the 9 tissue types. For each clas-



(a) Inputs (b) Expansions

Figure 2: Model inputs (a) and tile expansions (b) from the highest resolution of training (224×224px) for the CAMELYON17 dataset. Only the interior region in (a) are visible to the model.

sification task we trained a simple logistic model over training, validation, and test splits of each of the datasets. This model was implemented using the `scikit-learn` [29] library setting the maximum iterations to 1500 to ensure convergence. We report the performance of our model on the CAMELYON17 test dataset using accuracy, precision, recall and F1-score while we report overall balanced accuracy and F1-score for each of the classes of the CRC test dataset.

As baseline feature extractors, we compare our method to another popular self-supervised algorithm, SimCLR [6], as well as a pre-trained ResNet18 network with ImageNet weights. SimCLR minimizes a contrastive loss between encodings of similar patches and maximizes this loss between encodings of different ones. The contrastive loss is computed on mini-batches of paired images, which are strategically augmented in different ways. In this way, one enforces the encoder to find features that better equip the model to distinguish between different patches. For the training of SimCLR, the applied augmentations were color jitter, random rotation and horizontal flipping, color dropping, HED augmentation [34], image cutout [11], Gaussian noise, Gaussian blur and random cropping. The model was trained using SGD with Nesterov momentum and a batch size of 256 for 100 epochs. Two 256-dimensional fully-connected layers were appended to the 512-dimensional global pooling of the ResNet18 backbone. Note that these dense layers were discarded after training as in [6].

Representation	Accuracy	Precision	Recall	F1-score
Supervised †	**90.88%**	92.10%	**89.43%**	**90.75%**
Supervised	85.26%	**94.00%**	75.33%	83.64%
ImageNet	84.08%	84.56%	**83.39%**	83.97%
SimCLR[6]	84.61%	86.69%	81.78%	84.16%
Proposed w/o extrapolation	85.40%	88.74%	81.08%	84.74%
Proposed w/o D_Z	84.12%	**88.90%**	77.96%	83.07%
Proposed	**85.69%**	88.32%	82.25%	**85.18%**

Table 2: Performance of the different models for the test set of the CAMELYON17 dataset. A logistic regression model was used for all feature representations except Supervised † for which the performance was calculated using the ResNet18 [13] softmax output. With bold we highlight the best supervised and self-supervised performances.

As an additional baseline we trained completely supervised models using the same ResNet18 backbone network initialised with identical ImageNet pretrained weights. For these experiments, we appended a final dense classification layer with a softmax activation function. Furthermore, during training basic augmentation techniques were applied, such as random flip, random contrast and HED augmentation [34]. All the supervised models were early stopped after approximately 110 epochs. The best model was trained with the AMSGrad variant of the Adam optimizer, categorical cross entropy loss, learning rate of 0.0001, and batch size of 64.

Representation	Balanced Accuracy	F1-score									
		ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM	Average
Supervised †	85.27%	83.94%	**99.30%**	82.52%	94.72%	**91.07%**	61.53%	85.49%	65.39%	91.93%	86.26%
Supervised	**86.26%**	**85.38%**	99.24%	**84.97%**	**95.11%**	89.78%	**67.77%**	**86.76%**	**66.17%**	**92.45%**	**87.27%**
ImageNet	79.00%	**91.58%**	99.05%	76.84%	87.82%	87.54%	60.50%	75.91%	42.73%	82.66%	82.27%
SimCLR [6]	76.29%	87.93%	**99.76%**	65.53%	90.23%	77.54%	59.20%	76.81%	39.62%	84.47%	80.03%
Proposed w/o extrapolation	75.67%	87.20%	98.60%	54.83%	91.62%	89.83%	55.49%	70.46%	43.70%	83.64%	80.28%
Proposed w/o D_Z	82.66%	88.37%	99.59%	77.26%	88.35%	89.97%	67.06%	**80.05%**	57.77%	85.57%	84.51%
Proposed	**85.11%**	88.58%	98.14%	**86.87%**	**91.86%**	**92.61%**	**68.64%**	79.89%	**61.06%**	**88.31%**	**86.30%**

Table 3: Accuracy and F1-score for the CRC dataset. A logistic regression model was used for all feature representations except Supervised † for which the performance was calculated using the ResNet18 [13] softmax output. F1-score is reported for the different tissue classes i.e., adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). With bold we highlight the best supervised and self-supervised performances.

We additionally performed an ablation study on the proposed model. In one variation, we removed the discriminator on the latent code, D_Z, allowing the latents $\mathbf{z}$ to assume an arbitrary and unconstrained distribution. In another variation, we replaced the self-supervised image expansion task with simple reconstruction of the input crop. Finally, we trained the model without progressive training, that is, initiating training at the highest resolution. In each case, all other architectural and training properties were identical.

As shown in Table 2, all the evaluated models report balanced overall accuracy above 84% for the CAMELYON17 dataset, with the Supervised † method exhibiting higher accuracy, recall, and F1-score than all other methods. However, our proposed framework reports similar performances, outperforming both the pretrained ImageNet and SimCLR methods in terms of accuracy (+1%) and F1-score (+1%). Indeed, the highly expressive representations that are generated from our image expansion model are depicted in Figure 3, in which we plot the 2D t-SNE embedding produced by the features of the proposed model for the two CAMELYON17 classes. We highlight the ground truth labels, metastatic and non-metastatic, in different colors. It can be observed that the two categories have been successfully separated.

In Table 2 we additionally present the performance of different variations of our method. From these ablation results we observe that both without the discriminator D_Z, and separately, without the image expansion task, the model produces less descriptive features for the downstream classification tasks. In the former ablation, we hypothesise that, in imposing a Gaussian prior on the latents, D_Z enacts a regularisation that prevents overfitting to the data. In the latter, we hypothesise that, in the absence of the self-supervised task, the model is less inclined to extract a rich, high-level representation that it can use to generate an image expansion. We forgo the evaluation of the model as trained without progressive training, as in this final ablation, the model tended swiftly towards mode collapse.

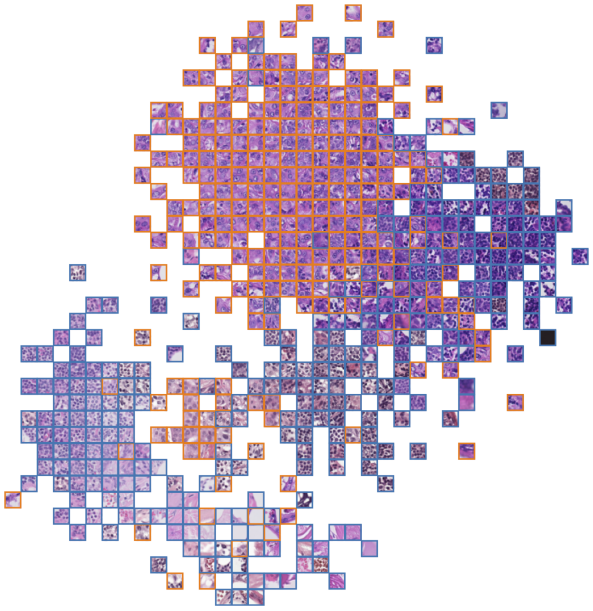


Figure 3: t-SNE plot of sample tiles from the test set used for the CAMELYON17 dataset. Metastatic tiles are outlined in orange, non-metastatic tiles outline in blue.

The overall accuracy and the performance per class in terms of F1-score for the CRC dataset is presented in Table 3. For these experiments, the supervised methods again outperform the others. However, our proposed method reports a competitive performance, outperforming the rest of the self-supervised methods by more than 6% in terms of overall accuracy. This highlights once more the expressive power of the representations generated by visual field expansion. In Figure 4 the t-SNE embedding for the CRC dataset is presented. The different classes are again highlighted with different colors, and we can observe a good

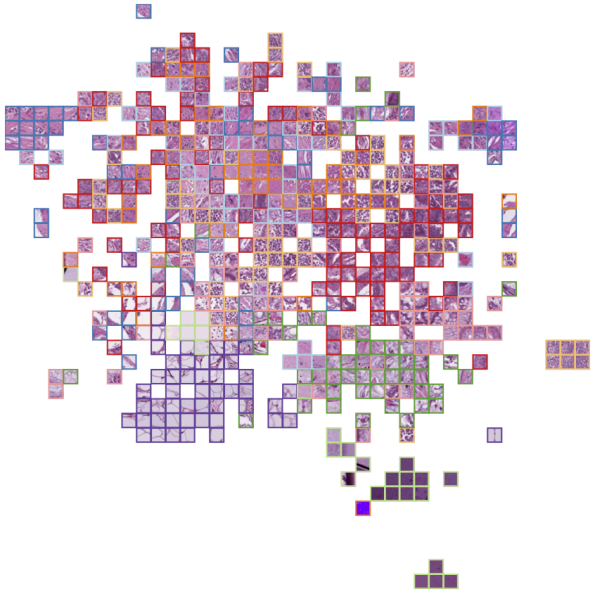


Figure 4: t-SNE plot of samples tiles from the test set used for the CRC dataset. Class is indicated by colour: STR (light blue), MUS (dark blue), BACK (light green), MUC (dark green), NORM (pink), TUM (red), LYM (light orange), DEB (dark orange), ADI (purple)

separation between the different categories. However, in this case we postulate the presence of regions of greater discontinuity in the image manifold. This may follow from the differing sampling policies of the two datasets: the CAMELYON17 tiles were sampled from continuously demarcated regions of tissue, while CRC the tiles are sampled from distinct tissue types. CRC may therefore lend more naturally to a non-Gaussian latent space.

With respect to the different classes that are presented, the smooth muscle (MUS) and cancer-associated stroma (STR) classes report the lowest performance for both supervised and self-supervised methods. Nevertheless, our proposed method reports a similar F1-score to that of the fully-supervised methods. For the remaining classes the F1-score is high for both supervised and self-supervised methods. Overall, in terms of average F1-score, our proposed method performs well and its performance is comparable to the fully supervised one. Once again, the proposed self-supervised formulation with both discriminators outperforms by a significant margin the ablation variants across the majority of classes, highlighting the gain in performance for our proposed design. Conversely, in a minority of classes, the performance of the proposed model is no better than the ablation models. We hypothesise that those classes, in particular background (BACK) and adipose (ADI), contain only sparse tissue, and defy the otherwise rich representations of the self-supervised model to find improvements.

6. Discussion

In this paper we proposed a generative model for extending the visual field of histopathology tiles. Our model is grown progressively, while two discriminators ensure a consistency among the fine details in the expanded regions and a structured latent space. The proposed method generates highly realistic images, preserving the structures and content that are presented in specific, predefined input patches. We perform extensive experiments on two publicly available datasets and report a FID of approximately 21 and 37 for CAMELYON17 and CRC datasets respectively for the expanded tiles. Moreover, our proposed framework simultaneously learns powerful representations that outperform commonly used pre-trained and self-supervised pipelines on two classification tasks, while reporting comparable performances to equivalent supervised methods. These promising results highlight the effective representation learning of the self-supervised task of visual field expansion.

One limitation of our method is that our framework has been trained such that the visual expansion of the content is performed from the center of the tile. This spatial prior constrains the generated tiles, without fully exploring the potentials of the problem of outpainting. Additionally, we have observed a tradeoff between good reconstruction and generation, with a divergence between the learned latent space and the target distribution. There is therefore scope for future work in reformulating the latent discriminator D_Z, or else replacing it with a divergence loss term. On the other hand, we have seen, in particular in the CRC dataset, evidence of non-Gaussianity of the latent space. Adversarial autoencoders are highly flexible models for specifying arbitrary template distributions for the latent space, and are therefore apt to exploring alternative formulations in the future.

Acknowledgements This work was supported by the ARC Grant SIGNIT201801286 and LabEx DIGICOSME scholarship (RD N^o 264). Finally we would like to thank Mihir Sahasrabudhe for all the constructive discussions during this project and Mesocenter[3] of CentraleSupélec for the computational resources.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster,

[3] http://mesocentre.centralesupelec.fr/

Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 4

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 3

[3] Aïcha BenTaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE transactions on medical imaging*, 37(3):792–802, 2017. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5, 6

[7] Ozan Ciga, Anne L Martel, and Tony Xu. Self supervised contrastive learning for digital histopathology. *arXiv preprint arXiv:2011.13971*, 2020. 2

[8] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. *Journal of Machine Learning for Biomedical Imaging*, 2021(4):1–48, 2021. 2, 4

[9] Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212*, 2018. 1

[10] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *International Conference on Medical Imaging with Deep Learning–Full Paper Track*, 2018. 2

[11] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 5

[12] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 6

[14] Ari Heljakka, Arno Solin, and Juho Kannala. Pioneer networks: Progressively growing generative autoencoder. In *Asian Conference on Computer Vision*, pages 22–38. Springer, 2018. 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[16] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2019. 2

[17] Le Hou, Vu Nguyen, Ariel B Kanevsky, Dimitris Samaras, Tahsin M Kurc, Tianhao Zhao, Rajarsi R Gupta, Yi Gao, Wenjin Chen, David Foran, et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern recognition*, 86:188–200, 2019. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 3, 4

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[22] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018. 3

[23] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021. 2

[24] Bin Li, Adib Keikhosravi, Agnes G Loeffler, and Kevin W Eliceiri. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. *Medical Image Analysis*, 68:101938, 2021. 2

[25] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 1, 3

[26] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. pages 1107–1110, 2009. 4

[27] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 3

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. 5

[30] Mihir Sahasrabudhe, Stergios Christodoulidis, Roberto Salgado, Stefan Michiels, Sherene Loi, Fabrice André, Nikos Paragios, and Maria Vakalopoulou. Self-supervised nuclei segmentation in histopathological images using attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 393–402. Springer, 2020. 2

[31] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1. 4

[32] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, page 101813, 2020. 2

[33] Dejan Štepec and Danijel Skočaj. Image synthesis as a pretext for unsupervised histopathological diagnosis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 174–183. Springer, 2020. 2

[34] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *CoRR*, abs/1902.06543, 2019. 5

[35] Maximilian E Tschuchnig, Gertie J Oostingh, and Michael Gadermayr. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns*, 1(6):100089, 2020. 2

[36] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021. 1

[37] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019. 2

[38] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ digital medicine*, 4(1):1–13, 2021. 1

[39] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10561–10570, 2019. 2

[40] Yushan Zheng, Bonan Jiang, Jun Shi, Haopeng Zhang, and Fengying Xie. Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 550–558. Springer, 2019. 1

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3