Medical Image Analysis 82 (2022) 102580

Contents lists available at ScienceDirect

Medical Image Analysis

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

Colour adaptive generative networks for stain normalisation of histopathology images

Cong Cong [a,*], Sidong Liu [b,c], Antonio Di Ieva [c], Maurice Pagnucco [a], Shlomo Berkovsky [b], Yang Song [a]

[a] School of Computer Science and Engineering, University of New South Wales, Australia
[b] Centre for Health Informatics, Macquarie University, Australia
[c] Computational NeuroSurgery Lab, Macquarie University, Sydney, Australia

ARTICLE INFO

Keywords:
Digital pathology
Stain normalisation
Generative adversarial networks,
Semi-supervised learning

ABSTRACT

Deep learning has shown its effectiveness in histopathology image analysis, such as pathology detection and classification. However, stain colour variation in Hematoxylin and Eosin (H&E) stained histopathology images poses challenges in effectively training deep learning-based algorithms. To alleviate this problem, stain normalisation methods have been proposed, with most of the recent methods utilising generative adversarial networks (GAN). However, these methods are either trained fully with paired images from the target domain (supervised) or with unpaired images (unsupervised), suffering from either large discrepancy between domains or risks of undertrained/overfitted models when only the target domain images are used for training.

In this paper, we introduce a colour adaptive generative network (CAGAN) for stain normalisation which combines both supervised learning from target domain and unsupervised learning from source domain. Specifically, we propose a dual-decoder generator and force consistency between their outputs thus introducing extra supervision which benefits from extra training with source domain images. Moreover, our model is immutable to stain colour variations due to the use of stain colour augmentation. We further implement histogram loss to ensure the processed images are coloured with the target domain colours regardless of their content differences. Extensive experiments on four public histopathology image datasets including TCGA-IDH, CAMELYON16, CAMELYON17 and BreakHis demonstrate that our proposed method produces high quality stain normalised images which improve the performance of benchmark algorithms by 5% to 10% compared to baselines not using normalisation.

1. Introduction

Histopathology images provide valuable information of diseases and their effects on tissues (Gurcan et al., 2009). To assist with microscopic analysis of tissues and cells, staining is applied to highlight structural features and enhance their contrast. However, undesired effects during the staining procedure can potentially lead to variations in the staining appearance. For example, Fig. 1 shows the colour variations within the same dataset of histological samples. While pathologists can deal with such colour variations, the performance of machine learning-based algorithms in digital histopathology image analysis can be heavily affected.

A **target domain** in histopathology image datasets can be defined as a group of images with relatively homogeneous stain colours, whereas the rest of the images are considered to be the **source domain**. The purpose of stain normalisation methods is to normalise the source domain images and match their colour distribution to the target domain and hence to compensate for the negative impacts of stain colour heterogeneity within a dataset. Many studies have used stain normalisation as the first step in the pipeline of histopathology image analysis (Ciompi et al., 2017; Stanisavljevic et al., 2018; Gandomkar et al., 2018; Kumar et al., 2020). However, most of these methods have implemented traditional stain normalisation methods (Reinhard et al., 2001; Macenko et al., 2009) which match the stain colour with a selected template image. These methods are designed based on mathematical models but can be heavily biased if a less representative template image is selected.

Recently, generative adversarial networks ($GANs$) (Goodfellow et al., 2014) have been widely investigated in stain normalisation. Among these methods, cycle-consistent generative adversarial networks ($CycleGAN$) (Zhu et al., 2017) based approaches have been studied most widely. These methods perform stain normalisation without

* Corresponding author.
E-mail address: c.cong@student.unsw.edu.au (C. Cong).

https://doi.org/10.1016/j.media.2022.102580
Received 11 November 2021; Received in revised form 12 May 2022; Accepted 11 August 2022
Available online 27 August 2022

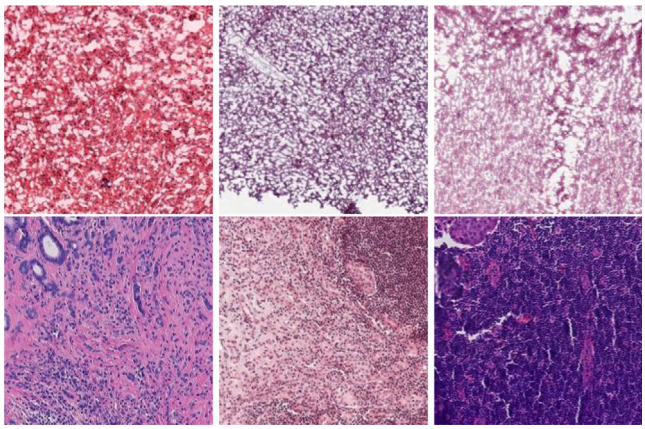


Fig. 1. Patch samples from TCGA-IDH (row one) and CAMELYON16 (row two). Colour heterogeneity can be observed within the same dataset.

requiring any template images and can generally achieve good results (Shaban et al., 2019). However, these methodologies still suffer from large discrepancies between source and target domains. In another study (Salehi and Chalechale, 2020), stain normalisation was treated as an image colourisation task, demonstrating improved performance over CycleGAN-based approaches. However, such methods require image colourisation outputs as ground truth for supervised learning and thus can only be trained on the target domain images. Such a setting does not directly represent the objective of stain normalisation, which is to normalise the colour of source domain images to that of the target domain images. To benefit from supervised image colourisation learning and also incorporate the source domain images into the colourisation learning process, we propose a colour adaptive generative adversarial network (CAGAN) for stain normalisation. Our method utilises the concept of *consistency regularisation* from semi-supervised learning so that source domain images can be used to enhance the learning of the colourisation model without requiring paired ground truth images.

1.1. Related work

1.1.1. Stain augmentation

Stain augmentation methods aim to reduce the model generalisation error by simulating variations in the dataset (Tellez et al., 2019). To mimic such variations, previous works conduct different types of image augmentations which can be roughly grouped into two categories: morphological and colour augmentations. Morphological augmentations simulate variations in morphological structures. Such augmentations typically include image rotation, flipping, elastic transformation, scaling and Gaussian blurring (Liu et al., 2017; Tellez et al., 2018a, 2019). Moreover, a wide range of colour augmentation methods have been proposed. These methods range from simple brightness, contrast and hue perturbations (Liu et al., 2017) to complicated operations on $H\&E$ colour space augmentation (Tellez et al., 2018a, 2019). Furthermore, recent studies have shown that using GAN (Wagner et al., 2021) or mix-up (Chang et al., 2021) to synthesise samples can further improve the model generalisation performance. These methods have shown that stain augmentations can be easily integrated with downstream models to enhance its robustness against stain variations, but they require prior knowledge about the data for careful design to guarantee effectiveness (refer to Section 4.1.1).

1.1.2. Stain normalisation

Stain normalisation approaches typically follow one of two categories: traditional methods and deep learning-based methods. Specifically, traditional methods use mathematical frameworks to match image features with a carefully selected template image. One group of studies focuses on aligning the colour distribution between two domains (Tabesh et al., 2007; Wang et al., 2007; Roy et al., 2019; Nadeem et al., 2020; Shafiei et al., 2020). These methods have been shown to produce satisfactory results but some (Roy et al., 2019; Nadeem et al., 2020) are at high computational costs for distribution estimation. For example, the computational complexity of these distribution alignment methods (measured in CPU time during inference) is normally 2∼3 times (Roy et al., 2019), in some extreme cases even 50 times (Nadeem et al., 2020) higher than the baselines (especially the traditional methods (Reinhard et al., 2001; Macenko et al., 2009). In fact, the high computational cost is a known limitation of these methods (Roy et al., 2019; Nadeem et al., 2020), making it hard to apply these methods in real-world clinical applications. Another group of studies sets out to normalise the stain vectors which are decomposed from the RGB space (Ruifrok and Johnston, 2001; Macenko et al., 2009; Li and Plataniotis, 2015b,a; Vahadane et al., 2016). These methods can be further improved by incorporating morphological features to produce cell-specific stain colour normalisation (Magee et al., 2009; Basavanhally and Madabhushi, 2013; Khan et al., 2014; Bejnordi et al., 2015; Janowczyk et al., 2017).

Deep learning-based stain normalisation methods usually treat stain normalisation as an image-to-image translation task and GAN is used often in these approaches. Specifically, **unsupervised deep stain normalisation** models are trained using both source and target domain images, wherein the source domain images are normalised to exhibit the target domain stain appearance, without requiring their paired input and ground truth images. Among these methods, $CycleGAN$ (Shaban et al., 2019) and its variants (de Bel et al., 2019; Shrivastava et al., 2019; Zhou et al., 2019; Mahapatra et al., 2020; Kang et al., 2020; de Bel et al., 2021) have been applied broadly in stain normalisation. Moreover, to save the effort of retraining a downstream task-specific network, BenTaieb and Hamarneh (2017) introduced a task-specific branch in the discriminator network and Nishar et al. (2020) combined an HRNet-based (Sun et al., 2019) generator with perceptual loss (Johnson et al., 2016) for better image content preservation. One drawback of unsupervised/unpaired stain normalisation methods is that their performance can deteriorate if the colour variations between domains is large. On the other hand, **supervised deep stain normalisation** methods (Cho et al., 2017; Zanjani et al., 2018; Tellez et al., 2019; Salehi and Chalechale, 2020; Cong et al., 2021a) are trained purely on target domain images. They normalise or colourise certain transformed representations (e.g., grayscale space) of the target domain images back to their original stain appearances (e.g., RGB space). While these approaches can produce high-quality normalisation results, their performance is constrained by the limited amount of target domain images available within a dataset. Moreover, the target-domain colourisation formulation does not fully resemble the objective of the stain normalisation between source and target domains.

1.1.3. Semi-supervised image colourisation

Inspired by semi-supervised learning (SSL) frameworks which combine unsupervised learning with supervised learning to obtain enhanced performance (Ouali et al., 2020a), in this paper, we focus on improving the previously proposed supervised deep learning-based methods by incorporating unsupervised learning on source domain images to enhance the performance of an image colourisation model. Specifically, we focus our work on the application of *consistency regularisation* (Lee et al., 2013; Tarvainen and Valpola, 2017; Park et al., 2018; Miyato et al., 2018; Verma et al., 2019; Berthelot et al., 2019; Ke et al., 2019; Sohn et al., 2020) and *proxy-labelling* (Lee et al., 2013; Yalniz et al., 2019; Arazo et al., 2020; Fang and Li, 2020; Ouali et al., 2020b; Xie et al., 2020), especially *co-training* (Han et al., 2018; Qiao et al., 2018) in our method.

Consistency regularisation, based on the assumption that a model's decision boundary should locate in the low-density region utilises unlabelled data to enhance the model by forcing it to produce consistent

outputs from an unperturbed unlabelled input and perturbed unlabelled input (Chapelle et al., 2009). While this simple and effective idea is widely used in semi-supervised deep learning for classification, the application of such a concept to image colourisation is not straightforward. As objects with similar shapes can have different colours, simply forcing the colourisation results to be consistent under perturbations can lead to a sub-optimal solution. To tackle this issue, the Transformation Consistency Regularisation (TCR) (Mustafa and Mantiuk, 2020) has recently been published. This work successfully incorporates consistency regularisation into image translation tasks by enforcing consistency between the prediction of a geometric transformation of an image and the geometric transformation of the prediction of the original image. However, based on our empirical studies, we find that such geometric transformation consistency is more helpful on natural image translation tasks, whereas consistency based on colour augmentations brings more benefits in a histopathology stain normalisation task.

Our work also explores the use of co-training (Blum and Mitchell, 1998). We design a model with a dual-decoder structure to enforce the consistency between two decoders as a form of self-training, wherein the outputs of the two decoders can be treated as perturbed views of each other. The benefits of this dual-decoder design are twofold: (1) it is a straightforward way to calculate the consistency loss between two decoders rather than designing a smoothing function to generate pseudo labels from multiple decoders; and (2) it is computational efficient as adding more decoders will increase the model complexity and require more computational resources to train.

1.2. Our contribution

In this paper, we propose a colour adaptive generative adversarial network (CAGAN) for stain normalisation to address the drawbacks of current supervised stain normalisation methods. Specifically, our contributions are summarised as follows:

- We propose a unified framework that combines both supervised learning from target domain and unsupervised learning from source domain.
- Our model adopts a novel dual-decoder design with consistency regularisation to enforce the generator to produce coherent stain normalisation results under perturbations.
- To properly adapt the concept of consistency regularisation to stain normalisation, we design two forms of perturbations: stain colour perturbation and model-embedded perturbation.
- Extensive experiments on the TCGA-IDH,[1] CAMELYON16[2]/17[3] and BreakHis[4] datasets demonstrate that our method improves the downstream classification task on three types of histopathology images.

Compared to our earlier work (Cong et al., 2021b), we apply stain colour augmentations as a source of perturbation on the inputs to simulate different situations which makes our colourisation model more robust to staining variations. Moreover, we improve the design of the loss functions with a histogram loss to generate better and more stable normalisation results. We have also conducted more extensive performance evaluation, ablation studies and included the additional CAMELYON16/17 datasets.

2. Materials & methods

2.1. Materials

We trained and evaluated our model as a preprocessing step of the tumour classification task using Hematoxylin and Eosin (H&E) stained histopathology images. Detailed information for each dataset is shown in Table 1.

Table 1
Overview of data used in this study.

Name	Slide count	Patch count	Tissue type
TCGA-IDH	1,494	–	Brain
CAMELYON16	400	–	Breast
CAMELYON17	100	–	Breast
BreakHis	–	7,909	Breast

For brain tumour classification, we focused on predicting the isocitrate dehydrogenase (IDH) gene mutation status as it is an important diagnostic, prognostic and therapeutic biomarker in glioma (Parsons et al., 2008). Specifically, we used the same dataset that we curated from The Cancer Genome Atlas (TCGA) program (Clark et al., 2013) in our previous study (TCGA-IDH) (Liu et al., 2020). TCGA-IDH consists of 1,494 whole-slide images (WSIs) from 921 glioma patients in the TCGA Lower Grade Glioma and Glioblastoma cohorts. Each patient has been labelled as either IDH wildtype (WT, n=517) or mutant (MU, n=404) based on immunohistochemistry and/or genetic sequencing.

Classification of breast cancer histopathology images into explicit histopathology patterns is of vital importance due to its high incidence rate in women (Siegel et al., 2021). In this work, we selected three breast cancer classification datasets for evaluation, including BreakHis, CAMELYON16 and CAMELYON17. The BreakHis dataset contains histopathology image patches of breast tumour tissue which were collected by surgical biopsy and labelled by pathologists from the P&D Lab (Spanhol et al., 2016). The aim of the task is to classify breast tumour tissue as benign or malignant. Specifically, surgical samples from 24 patients with benign breast tumours and 58 with malignant breast tumours were collected, forming a dataset of 2,480 benign and 5,429 malignant tumour images. The images contain a mixture of four magnification levels (i.e., 40X, 100X, 200X, and 400X).

We selected CAMELYON16 (Bejnordi et al., 2017) and CAMELYON17 (Bandi et al., 2019) for training and evaluating our proposed stain normalisation methods as a preprocessing step for breast cancer metastases classification in lymph nodes. In particular, CAMELYON16 contains a total of 400 WSIs from two medical centres, and CAMELYON17 contains 1,000 WSIs with 5 slides per patient collected from five medical centres. Following the experimental setup described in (Zhou et al., 2019; Mahapatra et al., 2020), we used CAMELYON16 for training the stain normalisation model and evaluate whether it improves the classification performance on CAMELYON17. The classification task on CAMELYON17 aims to discriminate between slides with metastases (positive) and without metastases (negative). We used the training set of CAMELYON17 for evaluation and the number of negative/positive slides of each of the 5 medical centres: 64/10, 58/10, 75/10, 60/10 and 61/10.

2.2. Methods

In this section, we first introduce the underlying GAN model for supervised stain normalisation on the target domain images. Then, we describe our modifications to facilitate colour adaptive learning incorporating source domain images. The overall model structure is shown in Fig. 2.

2.2.1. Problem definition

Given a histopathology image dataset I, we define a subset I_t which contains relatively homogeneous stain colours as the target domain and the rest of the images as the source domain I_s. The aim of stain normalisation is to reduce the stain colour discrepancy between the two domains, such that all images in I have the stain appearance of I_t. In this work, we use a pix2pix GAN (Isola et al., 2017) as our backbone network which requires paired image data to be trained. Ideally, these

[1] https://portal.gdc.cancer.gov/
[2] http://camelyon16.grand-challenge.org/
[3] http://camelyon17.grand-challenge.org/
[4] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

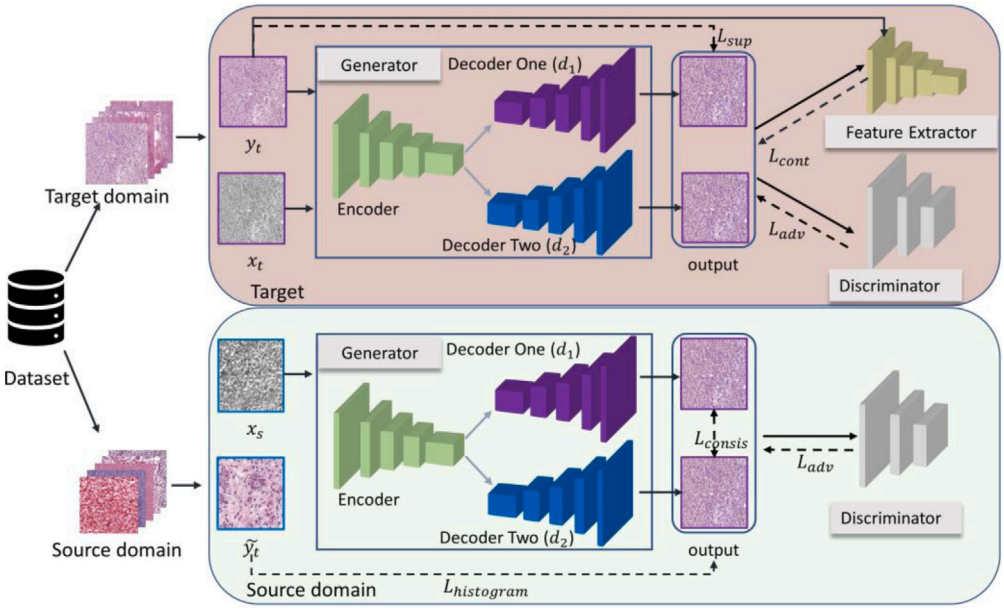


Fig. 2. The overall model structure. The generator takes the grayscale representations of the augmented patches from target and source domains as input $x_{t/s}$ and generates two normalisation results. This dual-decoder design is to perform unsupervised learning for source domain images in which two decoders can provide pseudo-labels for each other. Moreover, the introduction of content loss and histogram loss is to explicitly help the model generate image without losing structural and colour information.

pairs should be composed of images with different stain colours of the same sample. However, it is hard to obtain such image pairs in real-world datasets. Thus, we follow previous works (Cho et al., 2017; Salehi and Chalechale, 2020) and use the gray-scale transformation and the corresponding RGB image as paired image data. Specifically, given a transformed representation x_i, e.g., grayscale transformation of an image $i \in I$, we set out to train a model G_θ that colourises x_i with the stain colour of I_t. Furthermore, being aware that converting $H\&E$ stains into gray-scale images can cause information loss, we apply content loss (Johnson et al., 2016) and histogram loss (Afifi et al., 2021) to explicitly regularise the model to recover this information as much as possible and show further improved performance.

2.2.2. Supervised stain normalisation using target domain images

We first describe stain normalisation as a supervised image colourisation task using the target domain I_t only. Any target domain images $i_t \in I_t$, can be displayed in various representations. To train the colourisation model, we use its grayscale transformation x_t and its RGB representation y_t as labelled pairs (x_t, y_t). Specifically, we aim for the colourisation model to colourise x_t into y_t using a GAN model with supervised learning. We adapt a conditional generative adversarial network ($cGAN$) as our colourisation model. The generator G is used for colourising the inputs with the desired stain colours, whereas the discriminator D is designed to judge whether the colourisation results are from the target domain colour distribution or not. Here, x_t is used as input to G and we train G to map the inputs back to their original coloured appearance y_t, that is $G(x_t) = \hat{y}_t$, where $\hat{y}_t \approx y_t$. Unlike in the original GAN, discriminator D in $cGAN$ takes a paired input and assigns a higher value for actual RGB images y_t in the target domain and lower value for the colourisation results of G. In contrast, G tries to fool D by making D assign higher value for $\hat{y}_t$.

Network architecture. For the generator network (Fig. 4(a)), we used a U-Net (Ronneberger et al., 2015) structure with 5 down-sampling blocks *ConvBlock* and the same number of up-sampling blocks *UpConvBlock*. For each *ConvBlock*, we used batch normalisation followed by a convolutional layer and LeakyReLU is used as the activation function. Moreover, for each *UpConvBlock*, we applied *ConvBlock* to the up-sampled feature maps obtained using transposed convolution. For the discriminator network, a 5-layer PatchGAN (Isola et al., 2017) was used. Instead of producing a scale value, the PatchGAN network gives a $N \times N$ vector whose dimension depends on the shape of the input.

Regularisation. To train the colourisation model in a supervised fashion, we used the ground truth RGB y_t with shape $H \times W \times C$ as the label and apply $L1$ loss as the supervised loss L_{sup}:

$$\mathcal{L}_{sup} = \frac{1}{HWC} \left| \hat{y}_t - y_t \right| \tag{1}$$

We used adversarial loss $L_{adv_{D/G}}$ to update the discriminator and generator in turn, and we found that the least square loss from (Mao et al., 2017) can stabilise the training process as opposed to cross entropy in regular $GANs$.

$$\mathcal{L}_{adv_D} = (D(x_t, y_t) - 1)^2 + (D(x_t, \hat{y}_t))^2 \tag{2}$$

$$\mathcal{L}_{adv_G} = (D(x_t, \hat{y}_t) - 1)^2 \tag{3}$$

We further implemented the content loss L_{cont} (Johnson et al., 2016) to preserve the structural features as a form of content preservation. Specifically, a pretrained $VGG16$ model was used as a feature extractor. We take n layers of deep features for comparison. Since we extract features from both generated images $\hat{y}_t$ and the original RGB image y_t, this forms n pairs of feature maps. We then measured the distance between each feature map pairs to account for the content differences:

$$\mathcal{L}_{cont}(\hat{y}_t, y_t) = \sum_j^n \omega_j \frac{1}{C_j H_j W_j} ||\phi_j(\hat{y}_t) - \phi_j(y_t)|| \tag{4}$$

where ϕ_j is the feature map produced by the j_{th} layer before applying the max pooling operation, $C_j H_j W_j$ is the shape of ϕ_j and $\omega_j = 1/n$. In implementation, we use every layer before max-pooling to calculate content loss.

Thus, the total training loss for the supervised stain normalisation on the target domain is a combination of the above mentioned loss functions:

$$\mathcal{L}_{sup_D} = L_{adv_D} \tag{5}$$

$$\mathcal{L}_{G_{target}} = L_{sup} + L_{adv_G} + L_{cont} \tag{6}$$

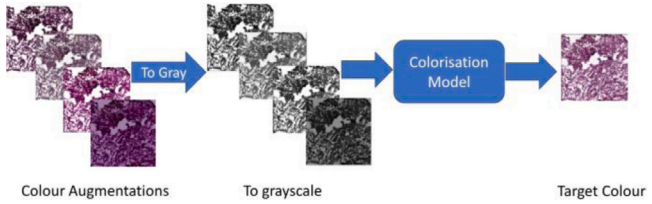


Fig. 3. Colour augmentations are applied to the inputs before they are introduced to the model which colourises them with the target domain colour appearances.

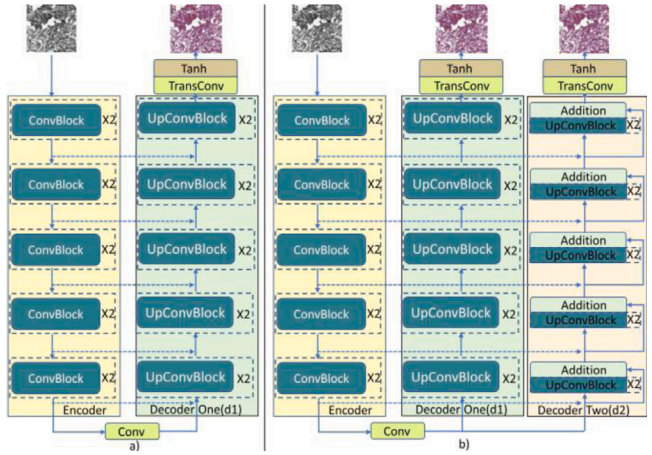


Fig. 4. Generator structures used in this work. The dotted arrow between encoders and decoders indicates the skip connection. Specifically, (a) shows the network in supervised stain normalisation; and, (b) shows the dual-decoder structure used in semi-supervised stain normalisation.

2.2.3. Colour adaptive learning using source domain images

In this section, we describe our modifications to the supervised stain normalisation GAN. Supervised stain normalisation models are trained on the target domain; however, they are not able to investigate the dataset fully as the source domain images are left unused. To make full use of the dataset, we incorporated the source domain images into the training of stain normalisation. However, this is not straightforward, as the colour appearance of source domain images y_s is not particularly useful in the supervised colourisation setting and the desired output after normalisation is not known. Therefore, inspired by the consistency regularisation from semi-supervised learning, we designed a colour adaptive learning framework which uses the source domain images by an adaptation of consistency regularisation. In particular, we designed a dual-decoder generator which outputs two colourisation results under some perturbations. Given an unlabelled source domain image as input, the generator produces two colourisation results which can be viewed as alternative representations of each other. By forcing consistency between them, we effectively assign a pseudo label for one decoder using the colourisation result from the other. In this way, the two decoders are able to learn from each other and the shared encoder can be further enhanced by the extra learning from the unlabelled source domain images. We also incorporate histogram loss (Afifi et al., 2021) to better regularise the colourisation results minimising the influence of semantic differences between the source and target domains.

Perturbations & Consistency Regularisation. The smoothness assumption in SSL classification states that, if two data points share the same label, their corresponding outputs should be the same. In the case of SSL image colourisation, we made a similar proposition that the colourisation output of an image should be consistent even if different transformations or different generators are applied to that image. Thus, in this work, we introduced two forms of perturbations, i.e., *input perturbation* and *model-embedded perturbation*, therefore asking the model to output consistent colourisation results under such perturbations. In particular, we applied input perturbations to simulate varying staining appearances. For this purpose, we used a combination of data augmentations, including Gaussian blur, contrast adjustment, saturation adjustment and H&E augmentation (Tellez et al., 2018b). Fig. 3 demonstrates some of the augmentation results, which show different staining colours of the same input image that should ideally be colourised with the same target domain colour appearance after stain normalisation. In addition, inspired by the concept of co-training illustrated in (Fang and Li, 2020; Ouali et al., 2020b), we present model-embedded perturbation, which generates perturbations via the design of model structures. Specifically, we modified the generator to have a dual-decoder structure and explicitly design these two decoder branches with different structures, as illustrated in Fig. 4. For *Decoder One* (d_1), we used the same structure described in Section 2.2.2, whereas for *Decoder Two* (d_2), we made a simple and effective modification which adds residual connections between the blocks. Such a design ensures the two decoders generate two alternative views of the same encoded feature and we show its effectiveness in Fig. 5 and Table 3. For any unlabelled source domain inputs, our generator outputs two colourisation results. Forcing the two results to be close to each other allows the model to be trained in an unsupervisedly manner by using the source domain images. Thus, for any source domain data x_s with shape $H \times W \times C$, the consistency regularisation L_{consis} can be formulated as minimising the mean absolute distance between two decoders' outputs:

$$\mathcal{L}_{consis} = \frac{1}{HWC}\left|G_{d1}(x_{s/t}) - G_{d2}(x_{s/t})\right| \tag{7}$$

Histogram Loss. It is likely that the two decoders generate consistent but incorrect colourisation results. A solution to this issue is to add an extra regularisation term which would align the output colour distribution with the target domain colour distribution. The recently proposed histogram loss (Afifi et al., 2021), which explicitly focuses on comparing the colour attributes was adapted into our method. To successfully incorporate histogram loss, we need to select a template image $\tilde{y}_t$ for the source domain images to match. Such a template should represent the target domain colour distribution. Here, we chose the target domain image whose pixel mean and standard deviation are closest to the overall mean and standard deviation of the target domain as our template. Then, given a source domain image x_s and template image $\tilde{y}_t$, we first converted them into log-chrominance space representations. Subsequently, we constructed histogram features H_s and $H_{\tilde{y}_t}$ by estimating the contribution of each pixel in the log-chrominance space to the histogram bins. To make the histogram feature differentiable for loss computation, we implemented an inverse-quadratic kernel κ with two tuneable parameters u and v to control the contribution of each pixel:

$$\kappa(I_{uc}, I_{vc}, u, v) = \frac{1}{1+(I_{uc}-\frac{u}{\tau})^2} \times \frac{1}{1+(I_{vc}-\frac{v}{\tau})^2} \tag{8}$$

where, I_{uc}, I_{vc} are the pixel intensity in log-chrominance space and τ is used to control smoothness. We apply this kernel function to each pixel in log-chrominance space to obtain the differentiable histogram feature H. Then we used the Hellinger distance to compute the histogram loss $L_{histogram}$:

$$\mathcal{L}_{histogram} = \frac{1}{\sqrt{2}}\left\|H_{\hat{t}}^{1/2} - H_s^{1/2}\right\| \tag{9}$$

Thus, we use the following equation to update the model using the source domain images where we replaced the content loss L_{cont} and supervised loss L_{sup} with histogram loss $L_{histogram}$ and consistency loss L_{consis}:

$$\mathcal{L}_{G_{source}} = L_{consis} + L_{histogram} \tag{10}$$

In summary, during the learning process, we use Eq. (5) to update discriminator and we update the generator using Eq. (6) for any inputs from the target domain and use Eq. (10) for the source domain inputs. Algorithm 1 shows the pseudo code of the overall training procedure.

Algorithm 1: Obtaining a trained generator of semi-supervised stain normalisation.

Data: m batches of target domain images pairs $\{(x_{ti}, y_{ti}) : i = 1, ..., m\}$, m batches of source domain images $\{x_{si} : i = 1, ..., m\}$, a pseudo label for the source domain image $\bar{y}_t$ and a random initialised model $p \ni (G_{\theta_g}, D_{\theta_d})$

Result: Trained models G_{θ_g} and D_{θ_d}

for *number of epoch* e **do**

  for *number of steps* k **do**

    $x_{ti} \leftarrow augment(x_{ti})$;

    $x_{si} \leftarrow augment(x_{si})$;

    Forward to obtain normalisation results:;

      $\hat{y}_{ti1}, \hat{y}_{ti2} = G_{\theta_g}(x_{ti})$;

      $\hat{y}_{si1}, \hat{y}_{si2} = G_{\theta_g}(x_{si})$;

    Updating discriminator:;

      $\nabla_{\theta_d} \frac{1}{m} \sum_1^m \mathcal{L}_{adv_D}(x_{ti}, y_{ti}, \hat{y}_{ti1/2})$;

    Updating generator:;

      $\mathcal{L}_{G_{target}} = \mathcal{L}_{adv_G}(x_{ti}, \hat{y}_{ti1/2}) + \mathcal{L}_{sup}(\hat{y}_{ti1/2}, y_{ti}) + \mathcal{L}_{cont}(\hat{y}_{ti1/2}, y_{ti})$;

      $\mathcal{L}_{G_{source}} = \mathcal{L}_{adv_G}(x_{si}, \hat{y}_{si1/2}) + + \mathcal{L}_{consis}(\hat{y}_{si1}, \hat{y}_{si2}) + \mathcal{L}_{histogram}(\hat{y}_{si1/2}, \tilde{x}_t)$;

      $\nabla_{\theta_g} \frac{1}{m} \sum_1^m (\mathcal{L}_{G_{target}} + \mathcal{L}_{G_{source}})$;

  end

end

Table 2
Train/test splits of each dataset.

Dataset	Train		Test	
	Slides	Patches	Slides	Patches
TCGA-IDH	1,191	17,686	149	2310
BreakHis-f1	–	5,005	–	2,904
BreakHis-f2	–	5,506	–	2,403
BreakHis-f3	–	5,332	–	2,577
BreakHis-f4	–	5,211	–	2,698
BreakHis-f5	–	4,826	–	3,083
CAMELYON16	270	319,861	129	120,129
CAMELYON17-C0	52	6,631	22	2,842
CAMELYON17-C1	48	7,312	20	3,134
CAMELYON17-C4	50	6,443	21	2761

3. Experiment

We first describe the datasets used in our experiments (details are show in Table 2). Then, we show the experiment setups which include both in-domain and cross-domain comparisons.

3.1. Dataset description

TCGA-IDH. TCGA-IDH (Liu et al., 2020) contains 1,494 slides, with 1,191 for training, 154 for validation and 149 for testing. Patches of size 1024 × 1024 pixels are cropped from each WSI at 10x magnification level and those with over 50% tissue contents are used. Since the images were collected from several tissue source sites (TSS), we selected training images which were collected from the largest TSS (contains most images) to form the target domain (200 slides) and the rest were used as source domain (991 slides).

BreakHis. For BreakHis, the provided images are of size 700 × 460 pixels. We mixed images of different magnification levels and conducted five-fold cross validation as published in (Spanhol et al., 2015). Furthermore, we used k-means (k=5) clustering to form clusters on training images based on pixel mean and standard deviation. Then, we selected the largest cluster as the target domain and the remaining images were treated as the source domain. Specifically, the distributions of images in target/source domain of each fold are: fold1: 299/4,706, fold2: 357/5,149, fold3: 324/5,008, fold4: 477/4734 and fold5: 337/4489.

CAMELYON16. CAMELYON16 contains 399 slides collected from two centres in which 270 slides are used for training and 129 slides are used for testing. We first loaded slides from 20× magnification level and then used the Otsu algorithm (Otsu, 1979) to filter out the background regions and randomly sampled image patches of size 256 × 256 from each slide. Then, we used the patches from the $Utrecht$ centre as the target domain (100 slides) and the patches from the $Radboud$ centre as the source domain (170 slides).

CAMELYON17. For each centre in the CAMELYON17, we mixed patches from 10 positive slides with those from negative slides and randomly selected 70% for training and 30% for testing. We used CAMELYON17 as an external evaluation dataset for cross-domain comparison (refer to Section 3.2) and we removed the two centres in CAMELYON17 which originally belong to CAMELYON16 and perform evaluation using the images from the other centres.

3.2. Experiment setup

To evaluate the quality of stain normalised images, we performed two sets of comparisons: **in-domain** comparison and **cross-domain** comparison. For **in-domain** comparison, we trained the stain normalisation models using the training set and evaluated the quality of stain normalised images on the test set separately for TCGA-IDH, BreakHis and CAMELYON16. To further examine the generalisation performance of CAGAN, we conducted **cross-domain** comparison in which we trained CAGAN on one dataset and evaluated its performance on the other datasets. Specifically, we measured the quality of stain normalised images on TCGA-IDH, BreakHis and CAMELYON17 using the trained model from CAMELYON16. For evaluation metrics, we use the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) (Wang et al., 2004) to measure the image quality of the stain normalised results. Both measures were obtained using the grayscale transformation of images before and after stain normalisation. Moreover, to evaluate colour consistency, we calculated the normalised median intensity (NMI). We used only the pixel intensity from the tissue region by filtering out the background regions using Otsu algorithm. Then we calculated the standard deviation NMI_{SD} and coefficient of variation NMI_{CV} as the measures of colour consistency.

We further evaluated the impact of stain normalisation on the classification performance. Specifically, we chose accuracy (Acc), F1-score ($F1$) and area under the receiver operating characteristic curve (AUC) as the metrics. Firstly, we conduct stain normalisation on all images in the dataset. Then, we trained a ResNet34 (He et al., 2016) classifier on the training sets and evaluate its performance on the test set. For BreakHis, we followed (Bayramoglu et al., 2016; Benhammou et al., 2020) to report mean and standard deviation of five-fold cross validation. To fairly compare with the baseline methods in TCGA-IDH and CAMELYON17, we used the train/test splits provided in Liu et al. (2020), Mahapatra et al. (2020) and reported results with 5 independent runs.

Both CAGAN and ResNet34 were developed using PyTorch on NVIDIA RTX 3090 GPUs. We obtained the best results by resizing the images to 256 × 256 for input and setting the batch size to 8. We used different learning rates for discriminator ($lr = 0.0003$) and generator ($lr = 0.0001$). We trained CAGAN for 100 epochs and ResNet50 for 40 epochs until convergence. Our code is available at https://github.com/thomascong121/CAGAN_Stain_Norm.

4. Results & discussion

In this section, we present our quantitative and qualitative evaluation results for both in-domain comparison and cross-domain comparison.

Table 3
SSIM and PSNR comparison with different methods on each dataset.

Method	TCGA-IDH		BreakHis		CAMELYON16	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Macenko	0.870 ± 0.035	23.41 ± 4.73	0.864 ± 0.024	23.92 ± 2.44	0.878 ± 0.117	22.31 ± 4.01
Reinhard	0.844 ± 0.044	23.64 ± 2.54	0.899 ± 0.022	26.27 ± 2.88	0.881 ± 0.176	20.20 ± 5.41
Vahadane	0.948 ± 0.037	26.14 ± 5.15	0.941 ± 0.010	28.75 ± 2.30	0.953 ± 0.017	28.29 ± 2.66
StainGAN	0.966 ± 0.024	28.45 ± 3.19	0.850 ± 0.021	23.93 ± 3.03	0.928 ± 0.028	19.98 ± 3.29
STST	0.912 ± 0.012	21.78 ± 5.60	0.935 ± 0.007	25.91 ± 1.35	0.973 ± 0.035	26.16 ± 4.87
Tellez et al.	0.965 ± 0.026	27.04 ± 3.42	0.916 ± 0.017	23.08 ± 0.78	0.961 ± 0.032	26.81 ± 2.32
Supervised	0.953 ± 0.011	26.06 ± 4.00	0.930 ± 0.011	26.85 ± 1.20	0.956 ± 0.031	25.58 ± 3.25
CAGAN	$\mathbf{0.984 \pm 0.013}$	$\mathbf{32.86 \pm 4.89}$	$\mathbf{0.951 \pm 0.006}$	$\mathbf{33.01 \pm 1.33}$	$\mathbf{0.986 \pm 0.014}$	$\mathbf{31.58 \pm 2.22}$

Table 4
NMI statistics comparison with different methods on each dataset.

Method	TCGA-IDH		BreakHis		CAMELYON16	
	NMI_{SD}	NMI_{CV}	NMI_{SD}	NMI_{CV}	NMI_{SD}	NMI_{CV}
w/o SN	0.046	0.057	0.050 ± 0.002	0.050 ± 0.003	0.065	0.067
Macenko	0.042 ± 0.008	0.051 ± 0.010	0.044 ± 0.006	0.051 ± 0.004	0.062 ± 0.005	0.065 ± 0.004
Reinhard	0.046 ± 0.004	0.052 ± 0.008	0.046 ± 0.003	0.052 ± 0.004	0.058 ± 0.004	0.060 ± 0.006
Vahadane	0.041 ± 0.002	0.045 ± 0.005	0.043 ± 0.008	0.048 ± 0.010	0.056 ± 0.012	0.068 ± 0.007
StainGAN	0.029 ± 0.004	0.034 ± 0.002	0.031 ± 0.007	0.038 ± 0.007	0.054 ± 0.003	0.064 ± 0.003
STST	$\mathbf{0.026 \pm 0.014}$	0.032 ± 0.010	0.035 ± 0.015	0.040 ± 0.018	0.050 ± 0.001	0.059 ± 0.002
Tellez et al.	0.028 ± 0.009	0.034 ± 0.011	0.037 ± 0.012	0.041 ± 0.014	0.052 ± 0.002	0.056 ± 0.002
Supervised	0.028 ± 0.012	0.036 ± 0.010	0.031 ± 0.012	0.036 ± 0.014	0.048 ± 0.002	0.057 ± 0.003
CAGAN	0.027 ± 0.006	$\mathbf{0.030 \pm 0.007}$	$\mathbf{0.030 \pm 0.013}$	$\mathbf{0.038 \pm 0.005}$	$\mathbf{0.040 \pm 0.002}$	$\mathbf{0.045 \pm 0.004}$

4.1. In-domain comparison

4.1.1. Quantitative comparison

Image quality of stain normalised images was evaluated using SSIM and PSNR as shown in Table 3. For both metrics, a larger value indicates better image quality. For comparison, we selected a range of stain normalisation benchmarks which include both traditional and deep learning-based approaches. In particular, for traditional stain normalisation approaches, we selected Macenko (Macenko et al., 2009), Reinhard (Reinhard et al., 2001) and Vahadane (Vahadane et al., 2016). For deep learning-based stain normalisations methodologies, we chose unsupervised methods (StainGAN (Shaban et al., 2019)) and supervised methods, such as stain-to-stain translation (STST) (Salehi and Chalechale, 2020), and the approach proposed by Tellez et al. (2019). Furthermore, we compared with the stain colour augmentation methods (StainAug) used in Tellez et al. (2018a, 2019) which uses both morphological augmentation and $H\&E$ augmentation. We also compared with using only the supervised part of our method as described in Section 2.2.2, which we refer to as "Supervised".

SSIM evaluates the structural feature similarity which can be used to measure the degree of structural preservation. CAGAN obtains the highest SSIM score over all the compared methods, indicating higher degree of structural preservation, which is partly attributed to the use of content loss L_{cont}. Moreover, incorporating unlabelled source domain images helps improve the overall stain normalised image quality. Table 3 shows that the PSNR score of CAGAN consistently outperforms other supervised and unsupervised methods. Additionally, we measured the stain colour consistency using NMI with different stain normalisation methods (the results are shown in Table 4). A smaller $NMI_{SD/CV}$ indicates better colour consistency. Generally, deep learning-based approaches generate stain normalised results with better colour consistency. Moreover, supervised methods tend to outperform the unsupervised methods, especially in the case where the colour variations between domains is large. Here, we use the difference between target domain NMI statistics and source domain NMI statistics ($Diff_{NMI} = \{Diff_{NMI_{SD}}, Diff_{NMI_{CV}}\}$) to measure the colour variations between domains. $Diff_{NMI}$ of TCGA-IDH, BreakHis, CAMELYON16 are $\{0.010, 0.009\}$, $\{0.008, 0.010\}$ and $\{0.018, 0.015\}$ respectively. Specifically, the $Diff_{NMI}$ of BreakHis is the average value of the five folds. Compared with supervised methods, unsupervised method generates stain normalised images with similar or even better colour consistency on TCGA-IDH and BreakHis, but it produces images with lower colour consistency on CAMELYON16. This indicates that, on a dataset which has large colour variations between domains, unsupervised stain normalisation methods tend to generate results with lower colour consistency. Moreover, CAGAN has the smaller standard deviation and coefficient of variation of NMI than most of the supervised methods, which shows the benefits of learning from extra unlabelled source domain images.

Table 5
Classification performance comparison of CAGAN and other stain normalisation methods on TCGA-IDH.

	TCGA-IDH		
	Acc	*F1*	*AUC*
w/o SN	0.837 ± 0.011	0.851 ± 0.012	0.878 ± 0.023
IDH_{study}	0.870	–	0.938
Macenko	0.867 ± 0.006	0.835 ± 0.015	0.909 ± 0.004
Reinhard	0.818 ± 0.003	0.830 ± 0.008	0.911 ± 0.007
Vahadane	0.897 ± 0.004	0.807 ± 0.005	0.925 ± 0.004
StainGAN	0.878 ± 0.006	0.873 ± 0.006	0.917 ± 0.006
STST	0.891 ± 0.004	0.880 ± 0.004	0.919 ± 0.002
Tellez et al.	0.916 ± 0.004	0.930 ± 0.004	0.948 ± 0.002
CAGAN	$\mathbf{0.941 \pm 0.005}$	$\mathbf{0.965 \pm 0.006}$	$\mathbf{0.983 \pm 0.005}$
StainAug	0.841 ± 0.003	0.865 ± 0.008	0.886 ± 0.015

Table 6
Classification performance comparison of CAGAN and other stain normalisation methods on BreakHis datasets.

	BreakHis		
	Acc	*F1*	*AUC*
w/o SN	0.825 ± 0.043	0.823 ± 0.066	0.910 ± 0.020
$BreakHis_{study1}$	0.890 ± 0.025	–	–
$BreakHis_{study2}$	0.834 ± 0.011	–	–
Macenko	0.938 ± 0.029	0.899 ± 0.044	0.885 ± 0.027
Reinhard	0.910 ± 0.034	0.911 ± 0.036	0.911 ± 0.024
Vahadane	0.908 ± 0.028	0.910 ± 0.036	0.921 ± 0.013
StainGAN	0.895 ± 0.015	0.811 ± 0.047	0.944 ± 0.012
STST	0.935 ± 0.024	0.924 ± 0.039	0.972 ± 0.017
Tellez et al.	0.939 ± 0.016	0.958 ± 0.044	0.969 ± 0.014
CAGAN	$\mathbf{0.981 \pm 0.004}$	$\mathbf{0.973 \pm 0.008}$	0.981 ± 0.014
StainAug	0.971 ± 0.027	0.957 ± 0.038	$\mathbf{0.984 \pm 0.016}$

Table 7
Cross-domain SSIM, PSNR and NMI statistics comparison of CAGAN (pretrained on CAMELYON16).

TCGA-IDH		BreakHis		CAMELYON17	
SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
0.970 ± 0.019	32.37 ± 3.86	0.973 ± 0.006	32.84 ± 2.85	0.935 ± 0.025	30.05 ± 1.98
NMI_{SD}	NMI_{CV}	NMI_{SD}	NMI_{CV}	NMI_{SD}	NMI_{CV}
0.028 ± 0.009	0.031 ± 0.007	0.025 ± 0.005	0.036 ± 0.008	0.046 ± 0.008	0.056 ± 0.007

Table 8
Cross-domain classification performance on TCGA-IDH and BreakHis of CAGAN (pretrained on CAMELYON116).

TCGA-IDH			BreakHis		
Acc	*F1*	*AUC*	*Acc*	*F1*	*AUC*
0.936 ± 0.009	0.946 ± 0.010	0.982 ± 0.005	0.968 ± 0.014	0.976 ± 0.011	0.985 ± 0.016

Tables 5 and 6 show the classification performance on stain normalised images processed by various stain normalisation methods. Specifically, for TCGA-IDH, we compared the results with the recent study (IDH_{study} (Liu et al., 2020)), which applies GAN for data augmentation aimed to improve the classification performance without stain normalisation. For BreakHis, we chose two deep learning-based approaches ($BreakHis_{study1}$ (Benhammou et al., 2020) and $BreakHis_{study2}$ (Bayramoglu et al., 2016)) that show good performance on the BreakHis dataset. Though the compared studies on TCGA-IDH and BreakHis datasets used exactly the same train/test splits as our study, Liu et al. (2020) used extra GAN-generated samples to obtain the best result. Since we do not have access to those synthetic samples, we only reported their best performance for comparison. It can be seen that our CAGAN achieves consistently higher performance on all datasets compared to the current state-of-the-art stain normalisation approaches. Specifically, CAGAN ranks first on TCGA-IDH and BreakHis in terms of Acc (0.944 ± 0.005; 0.981 ± 0.017), $F1$ (0.965 ± 0.006; 0.973 ± 0.026) and AUC (0.987 ± 0.005; 0.99 ± 0.005). Moreover, CAGAN shows slightly better performance than StainAug which was shown to outperform the stain normalisation methods in Tellez et al. (2019). However, we observed that the performance of StainAug dropped significantly on TCAG-IDH, which reveals that stain augmentation should be specifically designed for different datasets to work well.

Overall, stain normalisation, as a preprocessing step, improves classification performance to varying degrees ($1\% \sim 10\%$) across different datasets. Unlike deep learning-based stain normalisation methods, traditional approaches save the effort of training. Among the evaluated traditional approaches, Vahadane (Vahadane et al., 2016) outperforms the results by Macenko (Macenko et al., 2009) and Reinhard (Reinhard et al., 2001), producing images with a higher degree of structural preservation and less colour distortion but it also has higher computational complexity. Tables 5 and 6 show that Macenko (Macenko et al., 2009) and Reinhard (Reinhard et al., 2001) were able to increase classification performance in most cases. However, in some cases, their stain normalised images may decrease performance of a downstream classifier. This is due to the fact that template-based traditional stain normalisation methods overly rely on a single template image which may not be representative of the entire target domain.

In contrast, deep learning-based methods circumvent the heavy reliance on a template target image by directly learning the colour distribution of the target domain. All the presented deep learning-based stain normalisation methods can improve downstream classification performance. Moreover, by introducing semi-supervised learning on the source domain images, our CAGAN further improves on the supervised stain normalisation methods.

4.1.2. Qualitative comparison

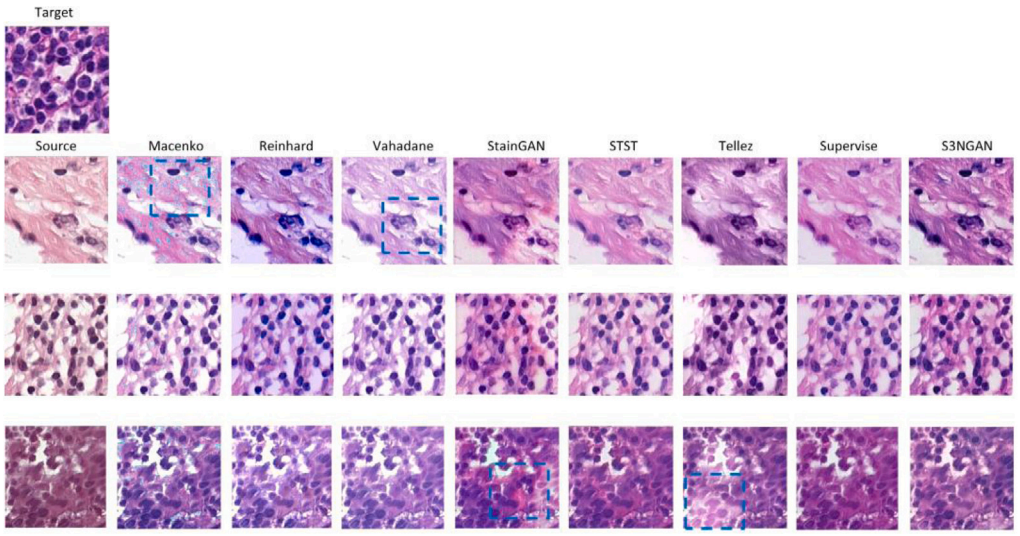
For qualitative results, Fig. 5 presents the normalisation results of three datasets using different stain normalisation methods. From the results, it is possible to observe that traditional stain normalisation approaches, such as the ones proposed by Macenko (Macenko et al., 2009) and Reinhard (Reinhard et al., 2001), are likely to generate outputs with apparent artifacts. In contrast, the method by Vahadane et al. (Vahadane et al., 2016) produces stain normalised images with much better quality, however, the nuclei are shown with lower contrast. On the other hand, it can be observed from Fig. 5 that the unsupervised CycleGAN-based approach (StainGAN (Shaban et al., 2019)) generates images with lower quality possibly due to the large discrepancy between the two domains, whereas supervised stain normalisation methods can produce images which better inherit the colours from the target domain. However, colour inconsistency and colour artifacts still persist in the results of these supervised approaches. This may be due to the fact that purely supervised-learning on the target domain cannot fully exploit the semantic features of the whole dataset which leads to parts of the nuclei being colourised incorrectly. Our CAGAN incorporates semi-supervised learning on the source domain images and generates stain normalised images which not only preserve image content of the source images, but also show better contrast of cell structures.

4.2. Cross-domain comparison

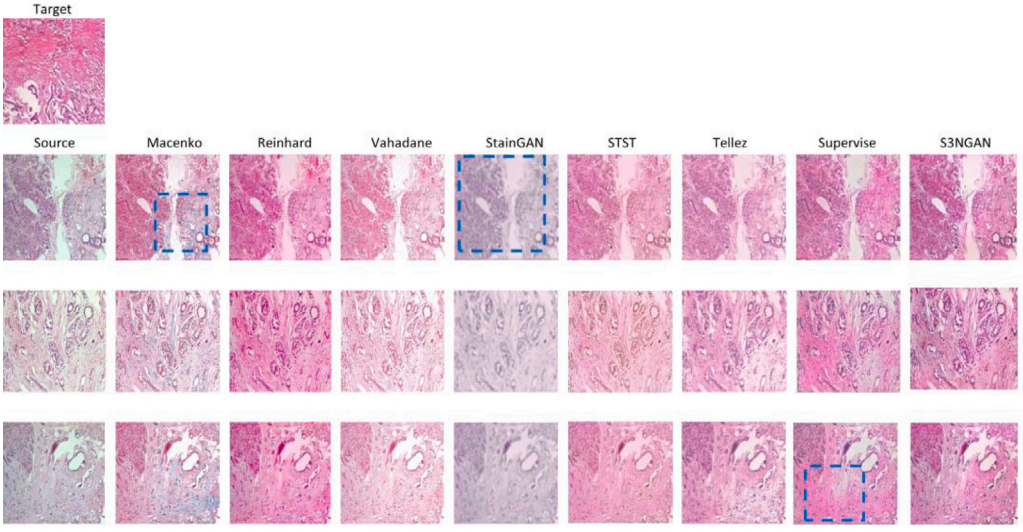
4.2.1. Quantitative comparison

We define $CAGAN_{cross}$ as the model trained using the cross-domain setting and $CAGAN_{in}$ as the model trained using the in-domain setting. In terms of image quality, we show the value of SSIM and PSNR of the images generated by $CAGAN_{cross}$ in Table 7. From the table, we notice that the mean values of PSNR and SSIM of $CAGAN_{cross}$ are close to those of $CAGAN_{in}$ on TCAG-IDH and BreakHis. Furthermore, we measured the colour consistency in terms of NMI. As shown in Table 7, interestingly, $CAGAN_{cross}$ actually generates images with higher colour consistency on BreakHis compared to $CAGAN_{in}$. Specifically, the mean $NMI_{SD/CV}$ of $CAGAN_{cross}$ are $0.002 \sim 0.003$ lower than value of $NMI_{SD/CV}$ in $CAGAN_{in}$. These results indicate the robustness of our proposed CAGAN.

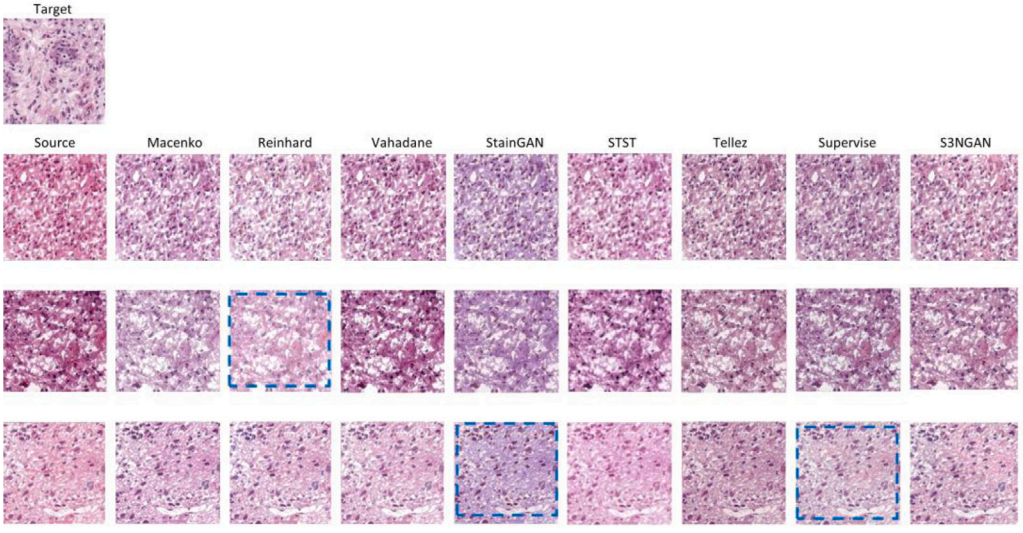
The effects of $CAGAN_{cross}$ on the downstream classification task are shown in Tables 8 and 9. While the classifier trained using the images generated by $CAGAN_{cross}$ performs equally well as $CAGAN_{in}$ on TCGA-IDH, it outperforms $CAGAN_{in}$ on BreakHis in terms of F1 score and AUC. Moreover, for CAMELYON17, we compared our results with the ones reported in two recent studies, the segmentation based colour normalisation network (SegCN-Net) (Mahapatra et al., 2020) and CNGAN (Zhou et al., 2019), which also trained stain normalisation on CAMELYON16 and conducted the testing on CAMELYON17. However, since we could not ascertain how prior studies distributed images across splits, the results from SegCN-Net and CNGAN are indicative of state-of-the-art performance, rather than offering direct result comparison. Overall, the classifier trained using the stain normalised images produced by CAGAN achieves consistently higher performance compared to the state-of-the-art approaches. These results show that our CAGAN is robust to domain changes and benefits from training on larger datasets.



(a) Stain normalisation results on the BreakHis dataset using different methods.



(b) Stain normalisation results on the TCGA-IDH dataset using various methods.



(c) Stain normalisation results on the CAMELYON16 dataset using various methods.

Fig. 5. Stain normalisation results on different datasets using various methods. The blue boxes indicate regions of some failure models, such as colour artifacts, incorrect colourisation, low contrast and colour inconsistency.

Table 9
Cross-domain classification performance on centre 0/1/4 in CAMELYON17 of CAGAN (pretrained on CAMELYON116).

		Acc	*F1*	*AUC*
C0	CAGAN	0.947 ± 0.002	0.948 ± 0.004	0.967 ± 0.015
	CNGAN	–	–	0.958
	SegCN-Net	–	–	0.967
C1	CAGAN	0.951 ± 0.005	0.956 ± 0.008	0.968 ± 0.012
	CNGAN	–	–	0.788
	SegCN-Net	–	–	0.864
C4	CAGAN	0.940 ± 0.012	0.947 ± 0.010	0.965 ± 0.017
	CNGAN	–	–	0.911
	SegCN-Net	–	–	0.946

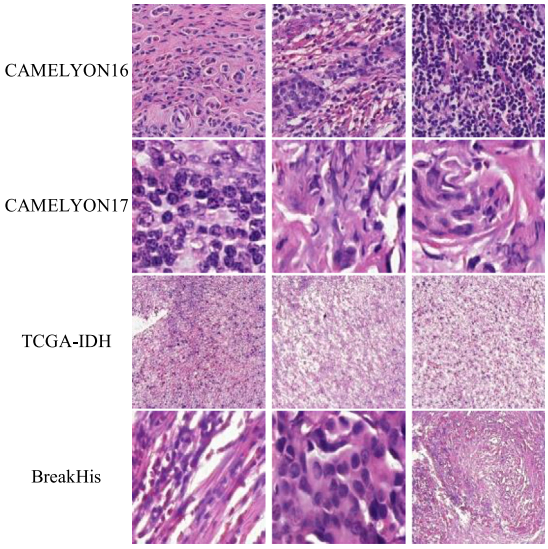


Fig. 6. Visualisation results of cross-domain evaluation. CAGAN was pre-trained on CAMELYON16 and evaluated on the other three datasets.

4.2.2. Qualitative comparison

Fig. 6 shows the results of cross-domain evaluation. CAGAN, which was pre-trained on CAMELYON16, is able to normalise images from other datasets into the CAMELYON16 stain colours without retraining the network. To further test if CAGAN is invariant to the change of magnification levels, we used various magnification levels for different datasets, among which BreakHis consists of image patches extracted from 40×, 100×, 200× and 400× and we used 10× for TCGA-IDH, 20× for CAMELYON16, 40× for CAMELYON17. As can be seen from the results, CAGAN which was trained on CAMELYON16 successfully normalises image from other datasets at various magnification levels.

4.3. Ablation studies

Loss Function Comparison. We also investigated the usefulness of applied histogram loss ($L_{histogram}$), content loss L_{cont} and adversarial loss. In particular, we evaluate the impact of (1) removing the histogram loss (w/o $L_{histogram}$); (2) removing the content loss (w/o L_{cont}); and (3) changing least-square to cross entropy for adversarial loss (L_{CE}). In the first plot of Figs. 8, 9 and 10, we notice that removing $L_{histogram}$ and L_{cont} from the regularisation terms always worsens the performance. As discussed in Section 2.2.3, the dual-decoder design of the generator helps the model learn from the source domain images by providing pseudo labels for each other. However, without applying $L_{histogram}$ as an extra constraint, they are likely to provide noisy labels. Moreover, we observe the largest performance drop without $L_{histogram}$. Fig. 7(d) shows one failure of the model without applying $L_{histogram}$ in which the normalisation result does not fully match the target domain's stain colour. This proves the importance of $L_{histogram}$ in helping the model to better align the resultant colourisation distribution with the target domain. Fig. 7(f) illustrates the role of L_{cont} in preserving the structural features of the source images. Specifically, we observe that the normalisation results do not represent the cell structures well and show less contrast without L_{cont}. However, implementing L_{cont} requires the use of a pretrained network which increases the computational overhead significantly. Future work may investigate other forms of content preservation to achieve similar effects with smaller computational overhead. With regard to the adversarial loss, in our previous work (Cong et al., 2021b), we show that using cross entropy is able to produce satisfactory stain normalisation results. In this work, we found that replacing cross entropy with least square loss leads to further improvement, as least square loss provides a smoother, non-saturated gradient for training the discriminator with accurate estimation of the distance from a data point to the decision boundary.

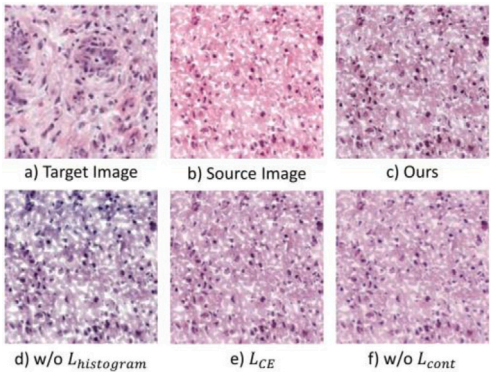


Fig. 7. Stain normalisation results on TCGA-IDH using different loss functions.

Input Perturbation Comparison. Input perturbation plays an important role in the success of consistency learning in unlabelled source domain images. In this work, we applied Gaussian blur, contrast adjustment, saturation adjustment and H&E augmentation (Tellez et al., 2018a) to simulate various stain variations. As can be seen in the second plot in Figs. 8, 9 and 10, the application of a combination of data augmentation leads to a 2%–7% improvement in terms of classification accuracy over the baseline with no augmentations. However, using a single augmentation might be insufficient. We observe that using Gaussian blur alone can sometimes negatively affect the model training especially on TCGA-IDH and CAMELYON17. As stated in (Tellez et al., 2019), Gaussian blur helps simulate out-of-focus defects due to improper use of scanners. We argue that such an augmentation benefits datasets which contain multi-resolution images such as BreakHis. This can be demonstrated from the results in Fig. 9 where using Gaussian blur alone can improve performance. It is noteworthy that we intentionally control the augmentation levels to produce in-distribution results which are more conservative.

Model-embedded Perturbation Comparison. Model-embedded perturbation is the main component which drives the model to learn in an unsupervised fashion from the unlabelled source domain images. As mentioned in Section 2.2.3, two decoder branches take the same encoded feature as input and generate two alternative views which are used as pseudo labels for each other. To make such a design effective, the implementation of Decoder Two ($d2$) should not decrease the performance of Decoder One ($d1$). In the third plot in Figs. 8, 9 and 10, we compared the performance with different choices of $d2$. Specifically, we firstly use the same structure as $d1$ but initialise it with different weights to obtain $d2$ and name this structure as $d2_{UNet}$. We observe unstable performance of $d2_{UNet}$ as random initialisation may lead to a network worse than $d1$. We then evaluated two choices of $d2$ that both have stronger learning abilities than a single multi-block

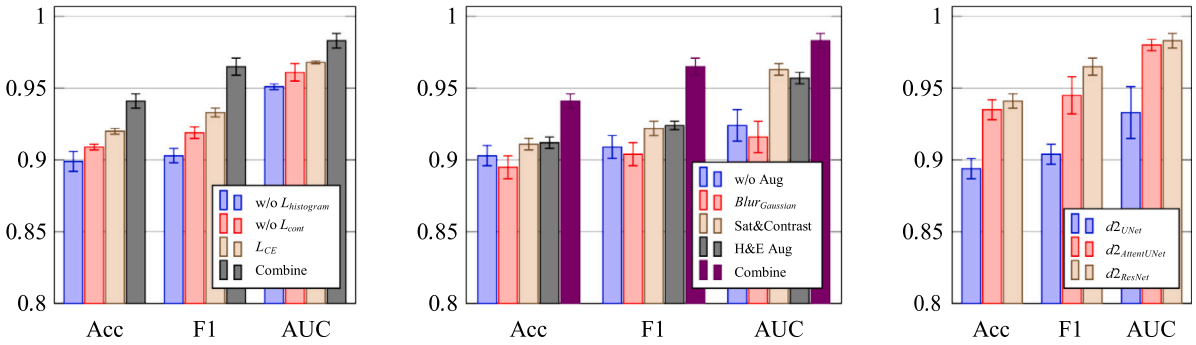


Fig. 8. Classification performance comparison with different losses (left), different augmentations (middle) and different design the second decoder (right) on TCGA-IDH.

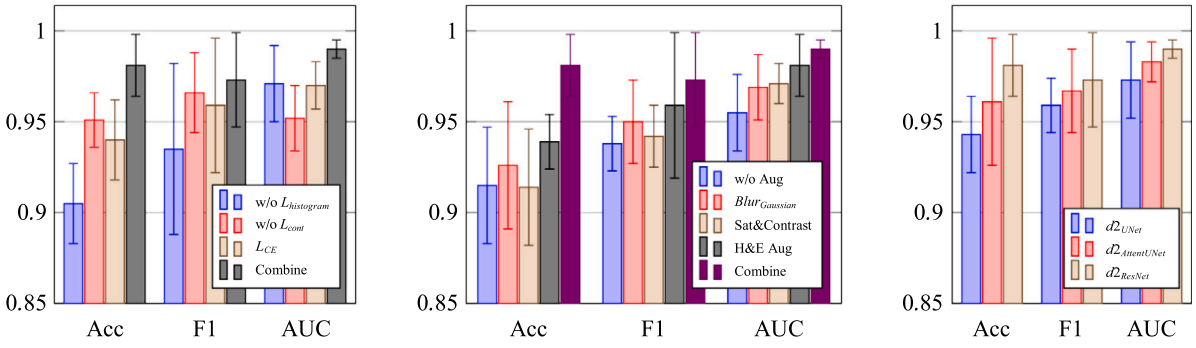


Fig. 9. Classification performance comparison with different losses (left), different augmentations (middle) and different design the second decoder (right) on BreakHis.

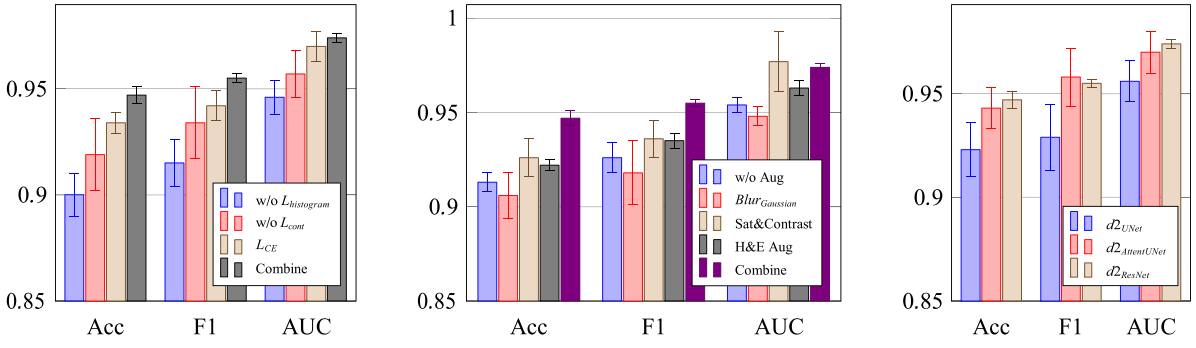


Fig. 10. Classification performance comparison with different losses (left), different augmentations (middle) and different design the second decoder (right) on CAMELYON17.

CNN ($d1$). They include 1) $d2_{AttentUNet}$ which incorporates channel-wise attention on skip connection at each level; and 2) $d2_{ResUNet}$ with residual connection between each convolution blocks (Fig. 4). Both designs obtain close performances, and $d2_{ResUNet}$ generates overall better stain normalisation results. Thus, we present $d2_{ResUNet}$ as our final design of $d2$.

Clinical Applicability. Our study, supported by extensive validation experiments, demonstrates that CAGAN is highly effective in stain normalisation and leads to improved performance in various histopathological image analysis tasks. For instance, with CAGAN, the IDH prediction model achieved an AUC of 0.987, compared to an AUC of 0.9 with no stain normalisation. As mutant IDH targeting therapies have been recently approved for the treatment of acute myeloid leukaemia (AML) (Issa and DiNardo, 2021) and as similar clinical trials for gliomas are already on-going (Chou et al., 2021), being able to accurately predict IDH mutation status from glioma H&E slides even before or without immunohistochemistry and/or genetic sequencing, which are the actual diagnostic gold-standard techniques, would allow more patients to benefit from precise treatments that target the underlying genetic cause of their cancer. The proposed method also improves the performance in benign and malignant breast tissue classification and breast cancer metastases classification in lymph nodes, thereby bringing it a step closer to clinical application in glioma and breast cancer diagnosis.

4.4. Limitations

Though we have shown the robustness of CAGAN, it is worth discussing its limitations. Firstly, CAGAN takes long time to train and requires high computational resources. Because of the introduction of an extra encoder, we were not able to use more than 8 images in a single batch on a NVIDIA RTX 3090 machine and it takes roughly about 5 to 6 days to finish the training on CAMELYON16. Moreover, we need to carefully select a reference target image for the histogram loss calculation. In future work, we aim to simplify the framework and design a reference image free loss function for the source domain colour regularisation.

5. Conclusion

In this work, we present a colour adaptive generative adversarial network (*CAGAN*) for stain normalisation which improves the current supervised stain normalisation approaches. We leverage the concepts of *co-training* and *consistency regularisation* from semi-supervised learning and design a dual-decoder generator in which each decoder outputs an alternative view of the same encoded feature. By forcing consistency between the two decoders, we allowed the model to effectively learn in an unsupervised manner from the unlabelled source domain images. Moreover, we introduced several colour augmentations as input perturbations which improve model robustness against stain colour variations. We extensively evaluated its effectiveness as an image pre-processing step on three histopathology image classification tasks. Results have shown that our method generates images exhibiting great consistency with the target domain which effectively improves downstream classification performance.

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.

References

Afifi, M., Brubaker, M.A., Brown, M.S., 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7941–7950.

Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al., 2019. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Med. Imaging 38 (2), 550–560.

Basavanhally, A., Madabhushi, A., 2013. EM-based segmentation-driven color standardization of digitized histopathology. In: Medical Imaging 2013: Digital Pathology, Vol. 8676. p. 86760G.

Bayramoglu, N., Kannala, J., Heikkilä, J., 2016. Deep learning for magnification independent breast cancer histopathology image classification. In: International Conference on Pattern Recognition. ICPR, pp. 2440–2445.

Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A., 2015. Stain specific standardization of whole-slide histopathological images. IEEE Trans. Med. Imaging 35 (2), 404–415.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Benhammou, Y., Achchab, B., Herrera, F., Tabik, S., 2020. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing 9–24.

BenTaieb, A., Hamarneh, G., 2017. Adversarial stain transfer for histopathology image analysis. IEEE Trans. Med. Imaging 37 (3), 792–802.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C., 2019. Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. pp. 92–100.

Chang, J.-R., Wu, M.-S., Yu, W.-H., Chen, C.-C., Yang, C.-K., Lin, Y.-Y., Yeh, C.-Y., 2021. Stain mix-up: Unsupervised domain generalization for histopathology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 117–126.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning (Chapelle, O. others, eds.; 2006)[book reviews]. IEEE Trans. Neural Netw. 20 (3), 542.

Cho, H., Lim, S., Choi, G., Min, H., 2017. Neural stain-style transfer learning using gan for histopathological images. arXiv preprint arXiv:1710.08543.

Chou, F.-J., Liu, Y., Lang, F., Yang, C., 2021. D-2-Hydroxyglutarate in glioma biology. Cells 10 (9), 2345.

Ciompi, F., Geessink, O., Bejnordi, B.E., De Souza, G.S., Baidoshvili, A., Litjens, G., Van Ginneken, B., Nagtegaal, I., Van Der Laak, J., 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, IEEE, pp. 160–163.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J. Digit. Imaging 26 (6), 1045–1057.

Cong, C., Liu, S., Di Ieva, A., Pagnucco, M., Berkovsky, S., Song, Y., 2021a. Texture enhanced generative adversarial network for stain normalisation in histopathology images. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1949–1952.

Cong, C., Liu, S., Ieva, A.D., Pagnucco, M., Berkovsky, S., Song, Y., 2021b. Semi-supervised adversarial learning for stain normalisation in histopathology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 581–591.

de Bel, T., Bokhorst, J.-M., van der Laak, J., Litjens, G., 2021. Residual cyclegan for robust domain transformation of histopathological tissue slides. Med. Image Anal. 70, 102004.

de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G., 2019. 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. pp. 151–163.

Fang, K., Li, W.-J., 2020. DMNet: Difference minimization network for semi-supervised segmentation in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI, Springer, pp. 532–541.

Gandomkar, Z., Brennan, P.C., Mello-Thoms, C., 2018. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif. Intell. Med. 88, 14–24.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. In: Conference on Neural Information Processing Systems. NIPS.

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.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M., 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. arXiv preprint arXiv:1804.06872.

He, K., Zhang, X., et al., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Isola, P., Zhu, J.Y., et al., 2017. Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134.

Issa, G.C., DiNardo, C.D., 2021. Acute myeloid leukemia with IDH1 and IDH2 mutations: 2021 treatment algorithm. Blood Cancer J. 11 (6), 1–7.

Janowczyk, A., Basavanhally, A., Madabhushi, A., 2017. Stain normalization using sparse autoencoders (StaNoSA): Application to digital pathology. Comput. Med. Imaging Graph. 57, 50–61.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. ECCV, pp. 694–711.

Kang, H., Luo, D., Feng, W., Hu, J., Zeng, S., Quan, T., Liu, X., 2020. Stainnet: A fast and robust stain normalization network. arXiv preprint arXiv:2012.12535.

Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W., 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6728–6736.

Khan, A.M., Rajpoot, N., Treanor, D., Magee, D., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans. Biomed. Eng. 61 (6), 1729–1738.

Kumar, A., Singh, S.K., Saxena, S., Lakshmanan, K., Sangaiah, A.K., Chauhan, H., Shrivastava, S., Singh, R.K., 2020. Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Inform. Sci. 508, 405–421.

Lee, D.-H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, Vol. 3, no. 2. ICML, p. 896.

Li, X., Plataniotis, K.N., 2015a. Circular mixture modeling of color distribution for blind stain separation in pathology images. IEEE J. Biomed. Health Inf. 21 (1), 150–161.

Li, X., Plataniotis, K.N., 2015b. A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. IEEE Trans. Biomed. Eng. 62 (7), 1862–1873.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442.

Liu, S., Shah, Z., Sav, A., Russo, C., Berkovsky, S., Qian, Y., Coiera, E., Di Ieva, A., 2020. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci. Rep. (Sci. Rep.) 10 (1), 1–11.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: IEEE International Symposium on Biomedical Imaging. ISBI, pp. 1107–1110.

Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K., Quirke, P., 2009. Colour normalisation in digital histopathology images. In: Proc Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy, Vol. 100. MICCAI Workshop, Citeseer, pp. 100–111.

Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Shao, L., 2020. Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 309–319.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802.

Miyato, T., Maeda, S.-i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1979–1993.

Mustafa, A., Mantiuk, R.K., 2020. Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, pp. 599–615.

Nadeem, S., Hollmann, T., Tannenbaum, A., 2020. Multimarginal wasserstein barycenter for stain normalization and augmentation. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 362–371.

Nishar, H., Chavanke, N., Singhal, N., 2020. Histopathological stain transfer using style transfer network with adversarial loss. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 330–340.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9 (1), 62–66.

Ouali, Y., Hudelot, C., Tami, M., 2020a. An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278.

Ouali, Y., Hudelot, C., Tami, M., 2020b. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684.

Park, S., Park, J., Shin, S.-J., Moon, I.-C., 2018. Adversarial dropout for supervised and semi-supervised learning. In: AAAI Conference on Artificial Intelligence.

Parsons, D.W., Jones, S., Zhang, X., Lin, et al., 2008. An integrated genomic analysis of human glioblastoma multiforme. Science 321 (5897), 1807–1812.

Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep co-training for semi-supervised image recognition. In: Proceedings of the European Conference on Computer Vision. Eccv, pp. 135–152.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Comput. Graph. Appl. (CG&A) 21 (5), 34–41.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention. MICCAI 2015, Springer International Publishing, Cham, pp. 234–241.

Roy, S., Lal, S., Kini, J.R., 2019. Novel color normalization method for hematoxylin & eosin stained histopathology images. IEEE Access 7, 28982–28998.

Ruifrok, A.C., Johnston, D.A., 2001. Quantification of histochemical staining by color deconvolution. Anal. Quant. Cytol. Histol. (AQCH) 23 (4), 291–299.

Salehi, P., Chalechale, A., 2020. Pix2Pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In: International Conference on Machine Vision and Image Processing. MVIP, pp. 1–7.

Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2019. StainGAN: Stain style transfer for digital histological images. In: International Symposium on Biomedical Imaging. ISBI, pp. 953–956.

Shafiei, S., Safarpoor, A., Jamalizadeh, A., Tizhoosh, H.R., 2020. Class-agnostic weighted normalization of staining in histopathology images using a spatially constrained mixture model. IEEE Trans. Med. Imaging 39 (11), 3355–3366.

Shrivastava, A., Adorno, W., Ehsan, L., Ali, S.A., Moore, S.R., Amadi, B.C., Kelly, P., Syed, S., Brown, D.E., 2019. Self-attentive adversarial stain normalization. In: International Conference on Pattern Recognition. ICPR.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer statistics, 2021. CA: Cancer J. Clin. 71 (1), 7–33.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685.

Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L., 2015. A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. (TBME) 63 (7), 1455–1462.

Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L., 2016. Breast cancer histopathological image classification using convolutional neural networks. In: 2016 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 2560–2567.

Stanisavljevic, M., Anghel, A., Papandreou, N., Andani, S., Pati, P., Hendrik Ruschoff, J., Wild, P., Gabrani, M., Pozidis, H., 2018. A fast and scalable pipeline for stain normalization of whole-slide images in histopathology. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5693–5703.

Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V.P., Verbel, D., Kotsianti, A., Saidi, O., 2007. Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE Trans. Med. Imaging 26 (10), 1366–1378.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Conferences on Neural Information Processing Systems. NIPS, pp. 1195–1204.

Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F., 2018a. H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In: Medical Imaging 2018: Digital Pathology, Vol. 10581. International Society for Optics and Photonics, p. 105810Z.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al., 2018b. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans. Med. Imaging 37 (9), 2126–2136.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging 35 (8), 1962–1971.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825.

Wagner, S.J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., Back, W.d., Peng, T., 2021. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 257–266.

Wang, Z., Bovik, A.C., et al., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, Y.-Y., Chang, S.-C., Wu, L.-W., Tsai, S.-T., Sun, Y.-N., 2007. A color-based approach for automated segmentation in tumor tissue classification. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 6576–6579.

Xie, Q., Luong, M.-T., Hovy, E., Le, Q.V., 2020. Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698.

Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D., 2019. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546.

Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A., et al., 2018. Histopathology stain-color normalization using deep generative models. In: International Conference on Medical Imaging with Deep Learning.

Zhou, N., Cai, D., Han, X., Yao, J., 2019. Enhanced cycle-consistent generative adversarial network for color normalization of H&E stained images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 694–702.

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. ICCV, pp. 2223–2232.