Medical Image Analysis 75 (2022) 102251

Contents lists available at ScienceDirect

Medical Image Analysis

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

High resolution histopathology image generation and segmentation through adversarial training

Wenyuan Li [a,b,*], Jiayun Li [a,c], Jennifer Polson [a,c], Zichen Wang [a,c], William Speier [a,c,d], Corey Arnold [a,b,c,d,e,**]

[a] Computational Diagnostics Lab, UCLA, Los Angeles, USA
[b] The Department of Electrical and Computer Engineering, UCLA, Los Angeles, USA
[c] The Department of Bioengineering, UCLA, Los Angeles, USA
[d] The Department of Radiological Sciences, UCLA, Los Angeles, USA
[e] The Department of Pathology & Laboratory Medicine, UCLA, Los Angeles, USA

ARTICLE INFO

Article history:
Received 5 October 2020
Revised 9 July 2021
Accepted 20 September 2021
Available online 3 November 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Histopathology image generation
Data augmentation
Semantic segmentation
Semi-supervised learning

ABSTRACT

Semantic segmentation of histopathology images can be a vital aspect of computer-aided diagnosis, and deep learning models have been effectively applied to this task with varying levels of success. However, their impact has been limited due to the small size of fully annotated datasets. Data augmentation is one avenue to address this limitation. Generative Adversarial Networks (GANs) have shown promise in this respect, but previous work has focused mostly on classification tasks applied to MR and CT images, both of which have lower resolution and scale than histopathology images. There is limited research that applies GANs as a data augmentation approach for large-scale image semantic segmentation, which requires high-quality image-mask pairs. In this work, we propose a multi-scale conditional GAN for high-resolution, large-scale histopathology image generation and segmentation. Our model consists of a pyramid of GAN structures, each responsible for generating and segmenting images at a different scale. Using semantic masks, the generative component of our model is able to synthesize histopathology images that are visually realistic. We demonstrate that these synthesized images along with their masks can be used to boost segmentation performance, especially in the semi-supervised scenario.

1. Introduction

Review of histopathology slides is important in medical diagnosis and treatment planning. It requires accurate quantitative analysis such as morphological feature extraction and cancer grading. Good segmentation may pave the way for these analyses and increase their reproducibility (Li et al., 2018). Recently, deep learning has brought significant improvement to semantic segmentation in medical image analysis. However, its performance typically relies on large annotated datasets (Litjens et al., 2017), and thus segmentation of histopathology images remains a challenge given the relative paucity of annotations. The images resulting from digitized histopathology slides are inherently high resolution (high-res), and obtaining large amounts of annotated data is laborious. Moreover, histopathological features in the images vary widely for different cancer grades, making them difficult to segment at a granular level.

Recently, generative adversarial networks (GANs) have been rapidly adopted by the medical imaging community (Yi et al., 2019). GANs have shown promise in data augmentation as they are able to synthesize high quality data to help overcome privacy issues and tackle the insufficiency of training data. However, most studies have focused on relatively low-resolution, small-scale images, such as CT and MRI (Armanious et al., 2020; You et al., 2019). The few GAN-based methods applied on pathology image synthesis focus only on cell-level feature representation (Kumar et al., 2017; Mahmood et al., 2019; Sirinukunwattana et al., 2016). High-res histopathology images contain diverse descriptors, and require GANs to preserve spatial consistency at a large scale, which has posed a challenge for their synthesis. Semi-supervised learning has emerged as a means to combine limited annotated data with abundant unannotated data to improve the training of ma-

* Corresponding author at: The Department of Electrical and Computer Engineering, UCLA, 924 Westwood Blvd Suit 420, Los Angeles, CA 90024, USA.
** Co-corresponding author.
E-mail addresses: liwenyuan.zju@gmail.com (W. Li), cwarnold@ucla.edu (C. Arnold).

https://doi.org/10.1016/j.media.2021.102251
1361-8415/

chine learning models. GANs have shown promising results in this regard. For example, Madani et al. (2018); Lahiri et al. (2017) and Lecouat et al. (2018) adopted the semi-supervised training scheme of GAN for chest abnormality classification, patch-based retinal vessel classification, and cardiac disease diagnosis, respectively. Yi et al. (2018) combined WGAN (Gulrajani et al., 2017) and CatGAN (Springenberg, 2015) for unsupervised and semi-supervised feature representation learning for dermoscopy images. Nevertheless, these methods have focused on classification tasks and cannot be directly applied to histopathology image segmentation. In this work, we propose a high-res large-scale histopathology image generation and segmentation framework through adversarial training. By setting up a dedicated multi-scale/pyramid training scheme, we are able to synthesize realistic histopathology images conditioned on semantic masks and use the synthesized images to train a segmentation network end-to-end in both fully-supervised and semi-supervised scenarios. Image synthesis for data augmentation using GAN is not new, but it is not widely used in histopathology analysis because generating images with fine details is difficult. Synthesizing images for gland segmentation poses even more challenges, as the generated images have to preserve both global gland structures and finer nuclear details based on the masks on a large scale. To the best of our knowledge, our approach is the first conditional generation model for high-res histopathology images and the first approach to use image synthesis for high-res histopathology image augmentation. Extensive experiments have been conducted to show the effectiveness of our proposed method in both image generation and semi-supervised segmentation. Detailed analysis is also provided to demonstrate how the multi-scale/pyramid structure and synthetic data augmentation each contribute to the model's performance.

2. Related work

2.1. Conditional GAN for image synthesis

Many researchers have leveraged conditional adversarial learning for image synthesis (also known as image-to-image translation), whose goal is to generate images based upon conditional input. For example, in natural images, pix2pix framework (Isola et al., 2017) used image-conditional GANs for different applications, such as generating cats from user sketches and transforming Google maps to satellite views. Subsequently, pix2pixHD (Wang et al., 2018a) proposed a multi-layer discriminator to synthesize high-resolution photo-realistic imagery without hand-crafted loss or pre-training. SinGAN (Shaham et al., 2019), on the other hand, introduced a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of a single image. In our proposed models, we incorporate the ideas of a multi-layer discriminator and construct the image generator in a pyramid fashion.

Medical images can also be generated by implementing constraints on segmentation maps. Guibas et al. and Costa et al. proposed a two-stage process that first trained a segmentation network to produce the vessel geometry, and then used the produced masks to synthesize fundus images (Guibas et al., 2017; Costa et al., 2017). Mok et al. proposed a coarse-to-fine network to generate brain MR images conditioned on a segmentation mask (Mok and Chung, 2018). Senaras et al. proposed a conditional GAN model for generating pathology images conditioned on nuclei segmentation masks (Senaras et al., 2018). Unlike the above models, our model works on high-res gland-level histopathology image generation, and we further leverage the synthetic images to train the segmentation tasks end-to-end in both supervised and semi-supervised settings.

2.2. GAN for segmentation

GANs have been used for segmentation tasks in medical images. In these cases, the discriminator can be regarded as a regulator and the adversarial loss can be viewed as a similarity measure between the segmented outputs and the annotated ground truth. Kamnitsas et al. proposed an unsupervised domain adaptation model using adversarial neural networks to train a segmentation task on brain MR datasets (Kamnitsas et al., 2017). Yang et al. achieved cross-modality domain adaptation, i.e. between CT and MRI images, via disentangled representations using adversarial training (Yang et al., 2019). Xue et al. used a multi-scale L_1 loss as a similarity measure in the BRATS challenges (Xue et al., 2018). Li et al. introduced an auxiliary classifier to regularize both the discriminator and the segmenter for fluorescent images (Li and Shen, 2018). Mahmood et al. demonstrates a nuclei segmentation method across different organs using deep adversarial training (Mahmood et al., 2019). Unlike the above models, our model consists of three components: a generator that can generate good images conditioned on the mask, a segmenter that can segment the input histopathological images, and a discriminator that distinguishes the ground-truth image-mask pairs from the pseudo image-mask pairs. The three network components form two adversarial games in training: one is between the generator and the discriminator that helps the generator to synthesize realistic images to compensate for the limited data size; the other is between the segmenter and the discriminator to help regularize the segmenter, so that the segmenter can output better masks to deceive the discriminator. Compared to the aforementioned models, our method provides two advantages for achieving better segmentation results: (1) the conditional generated images by the generator will be used to compensate the lack of training data (2) the adversarial game between segmenter and discriminator will regularize the model to learn the image-mask distribution.

2.3. GAN for semi-supervised learning

Several studies have adopted semi-supervised learning (SSL) training schemes using GANs in medical image classification problems. Madani et al. and Lecouat et al. found that an SSL-GAN can achieve comparable performance with traditional convolutional neural networks with less data in chest abnormality classification, retinal vessel classification, and cardiac disease diagnosis (Madani et al., 2018; Lecouat et al., 2018). Most of the other works that used GANs to generate new training samples applied a two stage process, with the first stage trained to augment the images and the second stage trained to perform a classification task. In contrast, our approach utilizes a single model that is capable of performing conditional synthesis and uses it to improve the downstream segmentation task simultaneously. Furthermore, there is limited research on segmentation in SSL on histopathology images. Zhang et al. proposed to use both annotated and unannotated images in a segmentation task, where the unannotated images are used to compute the segmentation masks to confuse the discriminator (Zhang et al., 2017). Bulten et al. used a semi-automatic segmentation method to generated semantic mask and grade prostate biopsies (Bulten et al., 2019). To the best of our knowledge, we are the first to explore GAN data augmentation effectiveness for segmentation in an SSL framework on histopathology images.

2.4. Semi-supervised segmentation for medical images

Semi-supervised learning is a fast-developing area of research. Besides GAN-based methods, there are other semi-supervised learning methods that have been developed by researchers such as co-training (Zhou et al., 2019; Xia et al., 2020) and methods that

use active learning (Bai et al., 2017). Among these methods, consistency regularization-based semi-supervised methods have been shown to be very effective. Teacher-student framework, a popular consistency regularization based-method, has been successfully applied to medical image segmentation tasks on several organs (Cao et al., 2020; Cui et al., 2019; Hang et al., 2020; Li et al., 2020; Sedai et al., 2019; Wang et al., 2020; Yu et al., 2019). In these approaches, the teacher model is a temporal ensemble of a current model with perturbations, while the student model learns from the teacher model through penalizing inconsistent prediction between the two, defined as consistency loss. One limitation of teacher-student models is that they depend heavily on the reliability of pseudo labels, and unreliable pseudo labels can have an adverse impact on training a segmentation network. A promising strategy to reduce this negative impact is to introduce model uncertainty. However, deriving model uncertainty measures is time consuming for large-scale high-resolution images. Moreover, consistency loss only penalizes independent pixel-level predictions, thereby not leveraging structure-level information in the learning procedure. Unlike the aforementioned methods, our proposed semi-supervised GAN model for high resolution histopathology images does not require time-consuming uncertainty measures. In addition, it incorporates a multi-layer generator and discriminator to learn information from different levels of structure, which could benefit segmentation tasks.

There are two main contributions of our paper. First, by using a pyramid generation scheme, we are able to effectively increase the receptive field of the segmenter and generate large-scale histopathological images up to 1024×1024 at high resolution ($20\times$). Compared to the state-of-the-art pathology synthesis methods, which generate images up to 256×256 allowing for limited context, such as nuclei (Mahmood et al., 2019; Senaras et al., 2018), our generation allows the incorporation of richer context, such as gland structures and nuclei details, that are useful for precise segmentation. Second, the generation is based upon a conditional method that produces high quality image-mask pairs. These image-mask pairs can be used to compensate for the lack of data points in training segmentation models. We demonstrate the effectiveness of our method in segmentation tasks and analyze how it performs in supervised and semi-supervised settings.

3. Methods

Our goal is to synthesize realistic histopathology images x based on an arbitrary semantic mask y, so that (x, y) can be used to compensate for a small data size when training a segmentation network. Subsequently, we show that the synthesized image-mask pairs can be used to boost segmentation performance, especially in the semi-supervised scenario seen in Section 4.

3.1. Generation

To synthesize high-res, large-scale histopathology images conditioned on semantic masks, our model must capture the statistics of complex image features at different scales. We wish to preserve global gland structures, such as shape and arrangement, while analyzing the finer details and textural information of the glands themselves, such as nuclei arrangement and lumen size. To achieve this, we propose to use a pyramid of conditional patch-GANs (Markovian discriminator) Isola et al. (2017).[1]

[1] Markovian discriminator, also known as PatchGAN, is a way to model high-frequencies components in images. It tries to classify if each $N * N$ patch in an image is real or fake.

3.1.1. Pyramid generation

Our framework consists of a pyramid of conditional generators, $\{G_0, G_1, \ldots, G_N\}$. It is trained on a pyramid of image-mask pairs (x, y): $\{(x_0, y_0), \ldots, (x_N, y_N)\}$ where (x_n, y_n) is a downsampled version of the original, (x_0, y_0). Each generator G_n attempts to generate realistic images $\tilde{x}_n$ conditioned on y_n. Through adversarial training, G_n learns to deceive an associated discriminator D_n, which attempts to distinguish (x_n, y_n) from $(\tilde{x}_n, y_n)$.

The pyramid framework begins at the coarsest scale G_N and proceeds sequentially to the finest scale G_0. Each G_n has the same architecture, and noise is injected at every scale to increase the variability among generated images. By progressing to finer scales throughout the generation process, the generators capture feature information of decreasing size. To start, G_N takes in a semantic map y_N with spatial white Gaussian noise z_N and maps it to an image $\tilde{x}_N$. At finer scales, G_n accepts an upsampled version of the generated image from the previous level $\tilde{x}^{\uparrow}_{n+1}$. The up-sampling is done via bi-linear interpolation. Spatial noise z_n is injected during this process, i.e. ,

$$\tilde{x}_n = G_n(\tilde{x}^{\uparrow}_{n+1}, y_n, z_n) \tag{1}$$
$$= \tilde{x}^{\uparrow}_{n+1} + \Phi_n(y_n, z_n + \tilde{x}^{\uparrow}_{n+1}). \tag{1}$$

Each of the generators G_n at finer scales ($n < N$) performs residual learning and adds details that are not generated by the previous scales, while maintaining features learned in previous steps of the pyramid. By going up in the generation process, finer details such as nuclei arrangement and lumen size are added while the global gland structures are preserved. Fig. 1(b) illustrates the details of G_n.

3.1.2. Multi-layer discriminator

Our discriminators take in image-mask pairs as input and differentiate whether they are real (x_n, y_n) or synthesized $(\tilde{x}_n, y_n)$ by the generator. To differentiate large-scale high-res real and synthetic images, the discriminator requires a large receptive field to stabilize training and improve generation performance. In practice, we found that using a multi-layer discriminator (Wang et al., 2018b) increases the training stability. Specifically, for each discriminator D_n, we downsample the real and synthetic image-mask pairs by factors of two and four to create a pyramid. Then the discriminators operate at each step of the pyramid to differentiate whether they are real or synthetic. The discriminators have identical architectures at each scale. Similar to the generators, their receptive fields get smaller at each finer scale. The discriminator at the coarsest view guides the generator to generate images that are globally, spatially consistent images, thereby preserving the gland structure based on semantic masks. The discriminator at the finest scale encourages the generator to produce finer details within this consistent structure. The multi-layer discriminator is illustrated in Fig. 1(d).

3.1.3. Training of generation

Our model is trained sequentially, from the coarsest scale to the finest scale. Once each scale is trained, it is kept fixed. Our training loss consists of four parts: adversarial loss, reconstruction loss, feature matching loss, and perceptual loss, i.e. ,

$$\min_{G_n} \max_{D_n} \quad \mathcal{L}_{adv}(G_n, D_n) + \alpha \mathcal{L}_{rec}(G_n) + \beta \mathcal{L}_{feat}(G_n) + \gamma \mathcal{L}_{perc}(G_n). \tag{2}$$

Adversarial loss The adversarial loss $\mathcal{L}_{adv}$ penalizes for the distance between the distribution of patches in (x_n, y_n) and the distribution of patches in the generated sample $(\tilde{x}_n, y_n)$ through a Markovian discriminator.

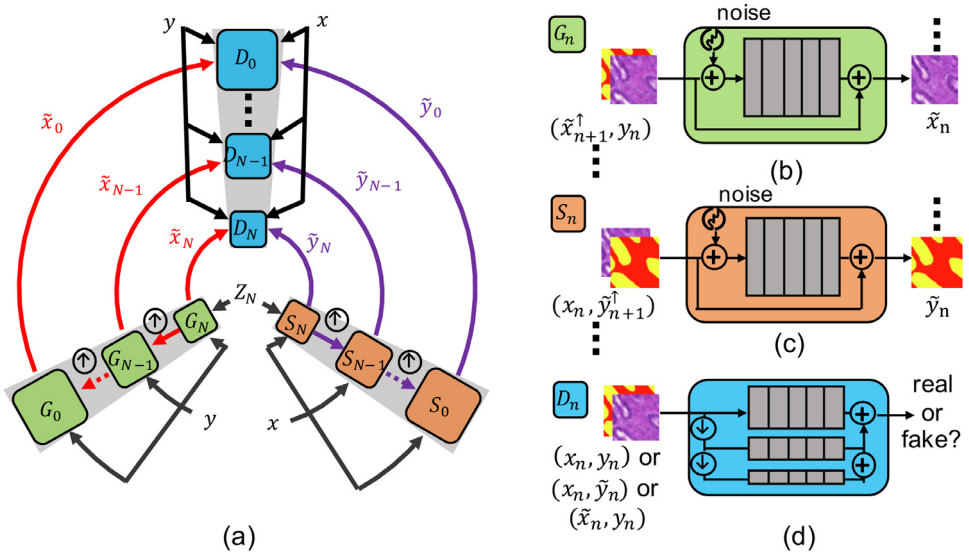

Fig. 1. Schematics of our approach. (a) A pyramid model that consists of a Generator G_n, a Segmenter S_n, and a Discriminator D_n at each scale. G_n synthesizes images based on mask y_n and lower-scale generation $\tilde{x}_{n+1}$; S_n segments images based on image x and lower-scale segmentation $\tilde{y}_{n+1}$; D_n enforces image-mask pairs from both G_n and S_n to match the real distribution. Once G_n achieves good results, we can use the synthetic data to train S_n. This path has been omitted in the figure for simplicity. Note that noise is injected to G_n and S_n, which has also been omitted in the figure too. (b) Illustration of the generator G_n. Each generator G_n attempts to generate realistic images $\tilde{x}_n$ conditioned on y_n and the previous generated images $\tilde{x}_{n+1}$. (c) Illustration of the segmenter S_n. S_n is symmetric with G_n as it conditions on the input image x_n and lower-scale segment results $\tilde{y}^{\uparrow}_{n+1}$, and attempts to segment the x_n. (d) Illustration of the discriminator D_n. D_n takes in image-mask pairs as input, and differentiates whether they are real (x_n, y_n) or fake $(\tilde{x}_n, y_n), (x_n, \tilde{y}_n)$ from the generator or the segmenter. To differentiate large-scale high-res real and synthetic images, we adopt a three-layer discriminator, which effectively increase the receptive field. $\uparrow, +$ denote upsampling and add operation respectively.

Reconstruction loss The reconstruction loss $\mathcal{L}_{rec}$ ensures that the generator is able to generate the original images based on the semantic mask.

$$\mathcal{L}_{rec}(G_n) = \left\| G_n(\tilde{x}^{\uparrow,rec}_{n+1}, y_n, z_n) - x_n \right\|^2. \tag{3}$$

Feature matching loss The feature matching loss $\mathcal{L}_{feat}$ is incorporated to improve the stability of training. We extract features from multiple layers of D_n and learn to match these intermediate representations from (x_n, y_n) and $(\tilde{x}_n, y_n)$ by L_1 loss.

$$\mathcal{L}_{feat}(G_n) = E_{(x_n,y_n)} \left\| D^{(i)}_n(x_n, y_n) - D^{(i)}_n(\tilde{x}_n, y_n) \right\|_1. \tag{4}$$

Perceptual loss The perceptual loss is also incorporated to ease the optimization (Wang et al., 2018b). We adopt a pre-trained VGGNet (Simonyan and Zisserman, 2014) for perceptual loss. Specifically, both the real and synthetic image-mask pairs are fed into a pre-trained VGGNet. We penalize the L_1 distance using features from the intermediate layers.

Note that $\mathcal{L}_{rec}$, $\mathcal{L}_{feat}$, and $\mathcal{L}_{perc}$ are only functions of G_n, i.e. we only use these losses to update G_n while keeping D_n fixed. We summarize the generation training algorithm in Algorithm 1.

Algorithm 1 Pyramid generation.

for Each scale of generation **do**
  for Number of training epochs **do**
    (1) Sample a batch of pairs $(\tilde{x}_n, y_n) \sim p_{G_n}(x_n, y_n)$ of size m_g, a batch of pairs $(x_n, y_n) \sim p(x_n, y_n)$ of size m_l;
    (2) Update D_n by ascending along its stochastic gradient based on Equation~(2);
    (3) Update G_n by descending along its stochastic gradient based on Equation~(2);
  end for
end for

3.2. Segmentation

To make the synthetic images useful for segmentation, we further design a pyramid structure with three players at each scale: (1) a segmenter S_n that characterizes the conditional distribution $p_{S_n}(\tilde{y}_n|x_n, \tilde{y}^{\uparrow}_{n+1}) \approx p(y_n|x_n)$, i.e. segmenting the input image based on the input image x_n and the semantic mask upsampled from the coarser level $\tilde{y}^{\uparrow}_{n+1}$; (2) a generator G_n that characterizes the conditional distribution in the other direction $p_{G_n}(\tilde{x}_n|y_n, \tilde{x}^{\uparrow}_{n+1}) \approx p(x_n|y_n)$, i.e. generating image based on the mask y_n and the upsampled synthetic image from the coarser level $\tilde{x}^{\uparrow}_{n+1}$; and (3) a discriminator D_n that distinguishes whether a pair of image-mask comes from the true distribution $p(x_n, y_n)$. While G_n and D_n are parameterized the same way as in Section 3.1, we use a mini FC-DenseNet (m-FCDenseNet) (Jégou et al., 2017) with 42 layers as S_n. The m-FCDenseNet was chosen because it has roughly 1 M trainable parameters, which is 17 times smaller than U-Net. With m-FCDenseNet, we are able to train the network for large-scale images with our GPU resources. The detailed architectures for G_n, S_n and D_n are shown in Appendix A.1. Fig. 1(c) illustrates the schematic of S_n. As mentioned above, S_n is symmetric with G_n since it takes in a upsampled version of lower-scale segment results $\tilde{y}^{\uparrow}_{n+1}$ with the image x_n.

$$\tilde{y}_n = S_n(x_n, \tilde{y}^{\uparrow}_{n+1}, z_n). \tag{5}$$

Accordingly, D_n, as an adversarial part of S_n, takes pseudo image-mask pair from segmenter $(x_n, \tilde{y}_n)$ and distinguishes it from the real distribution (x_n, y_n). The adversarial component between S_n and D_n can be formulated as a minimax game:

$$\begin{aligned} &\min_{S_n} \max_{D_n} \quad \mathcal{L}_{adv}(S_n, D_n) \\ &\mathcal{L}_{adv}(S_n, D_n) = E_{(x_n,y_n)}[\log(D_n(x_n, y_n)) \\ &+E_{(x_{S_n},y_{S_n})}[\log(1 - D(S_n(y_n|x_n, \tilde{y}^{\uparrow}_{n+1})))]. \end{aligned} \tag{6}$$

However, the game defined in Eq. (6) cannot guarantee that $p(x_n, y_n) = p(x_{s_n}, y_{s_n}) = p(x_{g_n}, y_{g_n})$ is the unique global optimum. To address this problem, we introduce the standard supervised

loss for segmentation (i.e. , cross-entropy loss) in S_n, $\mathcal{L}_{ce}(S_n) = E_{(x_n,y_n)}[\log(p_{S_n}(y_n|x_n,\hat{y}^{\uparrow}_{n+1}))]$. Consequently, the minimax game between S_n and D_n becomes:

$$\min_{S_n}\max_{D_n} \quad \mathcal{L}_{adv}(S_n,D_n)+\mathcal{L}_{ce}(S_n). \tag{7}$$

3.2.1. Training of segmentation

Training follows the same procedure as in Section 3.1.3, which starts from the coarsest scale and proceeds sequentially to the finest scale, except that for each scale we optimize D_n, G_n and S_n iteratively. Combining the minimax games defined in Eqs. (2) and (7), we formulate the game with three players G_n, S_n, D_n as:

$$\begin{aligned}\min_{G_n,S_n}\max_{D_n} \quad & \mathcal{L}_{adv}(G_n,S_n,D_n)+\alpha\mathcal{L}_{rec}(G_n)+\beta\mathcal{L}_{feat}(G_n)+\\ & \gamma\mathcal{L}_{perc}(G_n)+\alpha'\mathcal{L}_{ce}(S_n).\end{aligned} \tag{8}$$

The desired equilibrium of our model defined in Eq. (8) is that the joint distributions defined by the segmenter S_n and the generator G_n at each scale both converge to the true data distribution (Li et al., 2019). This is an important property, as pointed out by Li et al. , because it ensures that the generator G_n generates realistic image-mask pairs, enabling the segmenter S_n to leverage the synthetic image-mask pairs for training. We provide the detailed theoretical analysis of the equilibrium in Appendix A.2.

It should be noted that, during the initial stage of training at each scale, the synthetic images from the generator G_n are not realistic enough for training the segmenter S_n due to their low quality. Therefore, these generated image-mask pairs are not used to train S_n until the number of epochs reaches a threshold such that G_n can generate reliable image-mask pairs. In practice, we hold the synthetic images for 100 epochs and then use them as normal labeled image-mask pairs for training S_n, except the coefficient for the cross-entropy loss is smaller compared to the real annotated data (see details in Section 3.4). The threshold is determined by visually inspecting the synthetic images. Using the synthetic image-mask pairs in early training stages can disrupt the optimization process and potentially hurt segmentation performance. We summarize the segmentation training algorithm in Algorithm 2.

Algorithm 2 Pyramid generation and segmentation.

for Each scale of generation and segmentation **do**
  for Number of training epochs **do**
    (1) Sample a batch of pairs $(\tilde{x}_n,y_n)\sim p_{G_n}(x_n,y_n)$ of size m_g, a batch of pairs $(x_n,\tilde{y}_n)\sim p_{S_n}(x_n,y_n)$ of size m_s, and a batch of pairs $(x_n,y_n)\sim p(x_n,y_n)$ of size m_l;
    (2) Update D_n by ascending along its stochastic gradient based on Equation~(6);
    (3) Update G_n by descending along its stochastic gradient based on Equation~(6);
    (4) Update S_n by descending along its stochastic gradient based on Equation~(6);
  end for
end for

3.3. Semi-supervised segmentation

We further extend our framework to a semi-supervised learning (SSL) scenario. In SSL, we have a relatively small labeled set $(x_l,y_l)\sim p_l(x,y)$, and a large unlabeled set $x_u\sim p_u(x)$. We want to take advantage of the unlabeled data points x_u to boost our model's performance. To achieve this goal, we use the same architecture as in Section 3.2 with the following modification: at each scale, S_n takes in synthetic data, labeled data, and unlabeled data for training. We anticipate S_n to segment labeled data and synthetic data based on their masks (normal cross-entropy loss). For unlabeled images, we anticipate S_n to generate masks that are realistic enough that the image-mask pairs can confuse D_n through adversarial loss. The discriminator D_n accepts the image-mask pairs from the segmenter $S_n(x_{S_n},y_{S_n})\sim p(x_{u,n})p_{S_n}(y_{u,n}|x_{u,n})$, the generator $G_n(x_{G_n},y_{G_n})\sim p(y_n)p_{G_n}(x_n|y_n)$, and from the labeled data distribution $(x_{l,n},y_{l,n})\sim p_l(x_n,y_n)$ for judgement. D_n treats the labeled data as positive samples, and the pairs from both G_n and S_n as negative samples. By doing so, G_n and D_n, and S_n and D_n form two sets of adversarial training. The discriminator D_n will enforce both $S_n(x_{S_n},y_{S_n})\sim p(x_{u,n})p_{S_n}(y_{u,n}|x_{u,n})$ and $G_n(x_{G_n},y_{G_n})\sim p(y_n)p_{G_n}(x_n|y_n)$ to match with $(x_{l,n},y_{l,n})\sim p_l(x_n,y_n)$ during the training process, thus we will have a good generator G_n and good segmenter S_n at the end of the training.

One key problem in SSL is the limited amount of labeled data. A powerful D_n may memorize the labeled data and reject other types of samples. Consequently, G_n may collapse to these modes. To address this problem, we adopt the practical techniques introduced in (Chongxuan et al., 2017; Li et al., 2019). We generate pseudo masks through S_n for some unlabeled data and randomly choose these pairs as positive samples of D_n. This process introduces some bias to the target distribution of D_n, but it gives D_n a better chance to model the complete data distribution (Chongxuan et al., 2017; Li et al., 2019). Moreover, since S_n converges much faster compared to G_n, this operation enables G_n to explore a much larger image-mask distribution that includes both the labeled and unlabeled data information. In other words, S_n is able to provide pseudo masks for the unlabeled image x_u, while D_n will judge if the pseudo masks are reliable or not. This in turn will affect the evolution of G_n, which will take advantage of the unlabeled image to generate high quality images-mask pairs. These synthetic image-mask pairs that implicitly contain unlabeled data information will eventually benefit the training of the segmenter S_n. We will demonstrate that it serves as a key for performance improvement in SSL. We summarize the entire training procedure for SSL in Algorithm 3.

Algorithm 3 Pyramid semi-supervised segmentation.

for Each scale of generation **do**
  for Number of training epochs **do**
    (1) Sample a batch of pairs $(\tilde{x}_g,y_g)\sim p_{G_n}(x_n,y_n)$ of size m_g, a batch of labeled pairs $(x_l,y_l)\sim p(x_l,y_l)$ of size m_l, and a batch of unlabeled pairs $x_u\sim p(x_u)$ of size m_u;
    (2) Input x_u to S_n and get $(x_u,y_u)\sim p_{S_n}(x_n,y_n)$, input (x_l,y_l) to S_n and get $(x_l,\tilde{y}_l)\sim p_{S_n}(x_n,y_n)$;
    (3) Input $(x_l,y_l)\sim p(x_l,y_l)$, $(x_u,y_u)\sim p_{S_n}(x_n,y_n)$, $(x_l,\tilde{y}_l)\sim p_{S_n}(x_n,y_n)$, and $(\tilde{x}_g,y_g)\sim p_{G_n}(x_n,y_n)$ to D_n to get the output;
    (4) Update D_n by ascending along its stochastic gradient based on Equation~(8);
    (5) Update G_n by descending along its stochastic gradient based on Equation~(8);
    (6) Update S_n by descending along its stochastic gradient based on Equation~(8);
  end for
end for

Fig. 1 (a) illustrates our proposed pyramid model for histopathology image generation and segmentation.

3.4. Implementation details

In this work, we use Pytorch for training. We train the proposed model on a single Tesla V100S GPU with 32 GB memory.

We set the learning rates for G_n, S_n, and D_n to 1×10^{-4}, 5×10^{-4}, and 5×10^{-4}, respectively. We use an Adam optimizer and employ $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set the hyperparameters $\alpha, \beta, \gamma, \alpha'$ in Eq. (8) to be $0.001, 10, 10, 1$, respectively, same as in pix2pixHD and SinGAN (Wang et al., 2018b; Shaham et al., 2019). We do not further tune these hyperparameters, as they provide good generation and segmentation results. As mentioned above, during training, we first hold out the synthetic images for 100 epochs and then use them as normal labeled image-mask pairs for training the segmenter S_n. Once the synthetic data are used to train S_n, we set the coefficient of cross-entropy loss to 0.03 in order to decrease the adversarial effect of imperfect synthetic data. The whole network is then trained for 200 epochs. The number is empirically chosen by monitoring the loss function. We found that 200 epochs are enough for the network to converge. To choose the batchsize m_l, m_g, m_u in Algorithm 3, we follow the principle in Li et al. (2019). For the four-level training scheme, the batch size for each level is 8, 6, 4, 4 from the coarsest to the finest; for the five-level training scheme, the batch size for each level is 8, 6, 4, 4, 3 from the coarsest to the finest.

4. Experiments

In this section, we will first introduce the two datasets we used in our experiments, the GlaS and Prostate Gleason grading datasets. Then, we evaluate our proposed method in three aspects: image synthesis, image segmentation, and semi-supervised segmentation. We compare our method with baseline models, including pix2pix, pix2pixHD in image generation, mini FCDenseNet (mFCDenseNet), U-Net, DCAN etc. in segmentation, and demonstrate the effectiveness of our method under different scenarios (fully-supervised vs. semi-supervised).

4.1. Datasets

Our experiments are conducted on two histopathology image segmentation datasets, including the GlaS dataset (Sirinukunwattana et al., 2017) and the prostate Gleason Grading dataset (Gertych et al., 2015; Ing et al., 2018).

4.1.1. GlaS dataset

The GlaS dataset (Sirinukunwattana et al., 2017) was acquired by a team of pathologists at the University Hospitals Coventry and Warwickshire, UK. It consists of a training set with 85 images and a testing set with 80 images for colorectal cancer. The majority images are 775×522 pixel patches from whole-slide histology images of the colon. These images are scanned by Zeiss MIRAX MIDI and set to 20X magnification. The output was a color RGB image with the pixel size of 0.62 µm × 0.62 µm. Along with the images, pixel-wise annotations for epithelial glands (binary masks) and a spreadsheet detailing the type of the glands are provided. The types of glands are characterized as healthy, adenomatous, moderately differentiated, moderate-to-poorly differentiated, and poorly differentiated. The test dataset is divided into two subsets; subset A (60 images) released earlier and subset B (20 images) released during the original MICCAI workshop in 2015. We report results on the combined test set and the individual subsets.

4.1.2. Prostate Gleason grading dataset

The prostate Gleason grading dataset (Gertych et al., 2015; Ing et al., 2018) consists of 513 images. The dataset is retrieved from archives in the Pathology Department at Cedars-Sinai Medical Center (IRB# Pro00029960). The 513 images are combined from two sets of tiles. 224 of the images are from 20 patients and contain stroma (ST), benign or normal glands (BN, rated as GG2 or below), low-grade cancer (LG, image areas rated as GG3) and high-grade cancer (HG, image areas rated as GG4) (subset A). The remaining 289 images are from 20 different patients and contain dense high-grade tumors including Gleason grade 5 (GG5) as well as Gleason grade 4 (GG4) with cribriform and non-cribriform glands (subset B). Slides from subset A were digitized using a high resolution whole slide scanner SCN400F (Leica Biosystems, Buffalo Grove, IL), whereas slides from the subset B were acquired through the Aperio scanning system (Aperio ePathology Solutions, Vista, CA). The scanning objective in both systems was set to 20×. The output was a color RGB image with the pixel size of 0.5 µm × 0.5 µm and 8 bit intensity depth for each color channel. Representative tiles were extracted from whole slide images as 1200×1200 pixel tiles for analysis. The content of each tile was hand-annotated by an expert research pathologist using an in house developed graphical user interface. We use 80% of the images as training with the remaining 20% as testing unless otherwise specified. Note that both datasets are not split based on patients since a patient-level identifier is not provided with the data. Nevertheless, we believe our experimental comparisons are fair, as other work performed in the literature Li et al. (2017); Ing et al. (2018), and Li et al. (2018) for Prostate data; Chen et al. (2016) and Graham et al. (2019) for GlaS data) do not use patient-level splits.

4.1.3. Pre-processing

To handle different image sizes, we tiled the images into squares with overlap but without any scaling. Specifically, we tiled the GlaS images to 512×512, resulting approximately 500 patches, and the prostate images to 1024×1024, resulting approximately 1000 patches. The intensity value of the images was normalized to $[-1, 1]$. At training time we applied flipping, rotation, and color jitter to augment the data. When scaling, bi-linear interpolation was used for images while nearest neighbor method was used for masks. Fig. 2 shows some representative images of the cropped patches from both datasets.

4.2. Image generation

4.2.1. Qualitative evaluation

We first show the generation results of our model qualitatively. For the GlaS dataset, the model was trained on 512×512 image patches on two types of masks: binary masks (stroma vs. epithelial glands) and multi-category masks that indicated the gland type. Here, we made a minor assumption that all the glands in one single patch have the same type as indicated in the data spreadsheet. When training the model, we first downsampled the original images and started from patches of size 64×64. For each following scale, we multiplied the image length by a factor of two. Thus it led to a four-level training scheme with image size of $64^2, 128^2, 256^2, 512^2$. For the prostate Gleason grading dataset, the model was trained on 1024×1024 image patches starting from 64×64, which led to a five-level training scheme of $64^2, 128^2, 256^2, 512^2, 1024^2$. We compared our method with two baseline methods: pix2pix generation (Isola et al., 2017) and pix2pixHD generation (Wang et al., 2018b). To qualitatively analyze the results, we show samples of synthetic images in Fig. 3. The figure illustrates that our method preserves the global structure indicated by the semantic mask, and generates sharper images with finer details than the baseline methods. More synthetic high-resolution samples can be found in Appendix A.3.

Next, we demonstrate some interesting aspects during the model's generation process. Fig. 4(a–c) shows the generation results from the GlaS dataset conditioned on multi-category masks. Fig. 4(a) shows the coarsest-to-finest generation process. We observe that as the training process progresses, the generated images are honed (i.e. more details are added while the gland structures are preserved) so that the images look more realistic. Fig. 4(b)

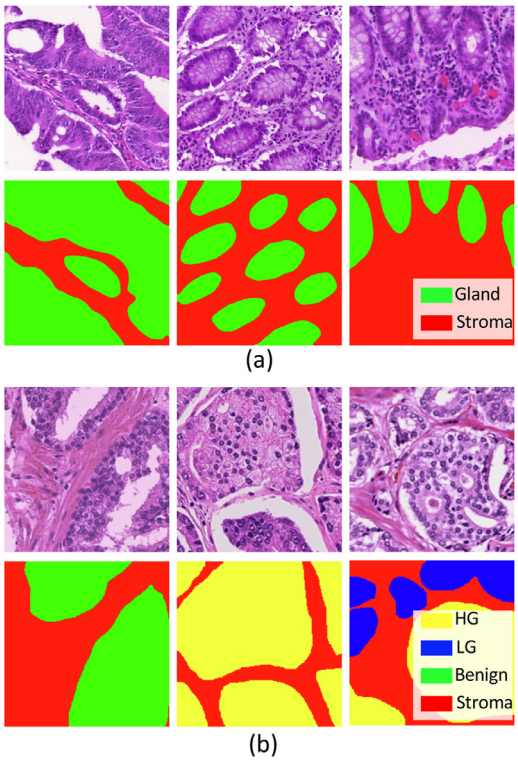

Fig. 2. Samples from the GlaS and prostate datasets. Three representative examples are shown from each dataset. (a) Samples from the GlaS dataset with their segmentation ground truth. Green color indicates the gland in the images while red color indicates stroma. (b) Samples from the Prostate dataset with their segmentation ground truth. Images are annotated by pathologists for stroma in red, benign glands in green, low-grade cancer in blue, and high-grade cancer in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

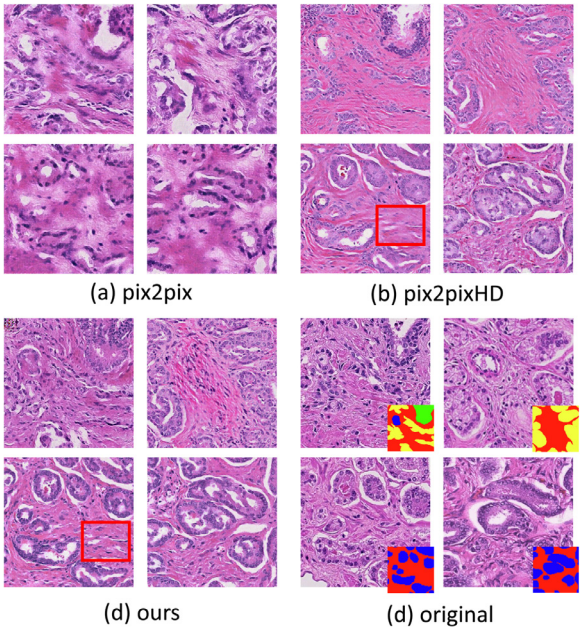

Fig. 3. Random synthetic images by different models and the original real images. The figures illustrate that our model preserves the global structures of the semantic masks and generates sharper images with finer details than the baselines. Pix2pix model cannot preserve the gland structure based on the semantic mask while pix2pixHD loses some finer details through the generation (indicated by the red box). Note that all images have the same field of view with the same pixel size (4e−4 μm/pixel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
FID score for image generation.

	GlaS	Prostate
pix2pix	3.654	4.354
pix2pixHD	1.028	2.577
Ours	0.0059	0.013
Original images	8.4×10^{-4}	2.3×10^{-4}

shows three images that are generated based upon the same mask. Once training was completed, we performed inference using the mask shown in the top-left corner. The generation process also started from 64×64 patch size and then went up to 512×512. Though all three of the images preserve global structures, subtle details are different due to the injected noise, e.g. the stroma details in the rectangle in Fig. 4(b). It can be more easily observed in the animation provided in the Appendix, where we cycle through these generated images. Since we injected noise during the generation process, we can continually generate images based upon the same mask and use them for training in the segmentation task. Fig. 4(c) shows image manipulation results by changing the mask from healthy to poorly differentiated. For these images, we changed the gland labels on the input masks and fed them into our generation framework to generate images of different grades with the same gland boundaries. It suggests that the generator can learn meaningful latent representations instead of simply memorizing the training data. Similar observations can be found for the Prostate dataset, where we changed the labels from low grade to high grade and vice versa. For more results on prostate dataset generation, noise injection, and image manipulation, we refer readers to Appendix A.3.

4.2.2. Quantitative evaluation

If our generated images are realistic looking, then their distribution should be indistinguishable from that of the real images. Therefore, we can quantitatively evaluate the quality of the synthetic images by computing the Frechet Inception Distance (FID) between the distributions of real images and synthetic images (Heusel et al., 2017). Lower values of FID indicate the distribution are more similar, implying more realistic-looking images. Specifically, we adopted a ResNext50 model pre-trained on large-scale histopathology images to extract features for computing FID. We didn't use the Inception v3 model that commonly used in other literature, as we could not find an off-the-shelf Inception model pre-trained on histopathology images. Therefore, our FID values are not directly comparable to common FID values in other studies. We provide the FID score of our model and other baselines in Table 1. As shown in Table 1, our model achieves lower FID compared to the baselines in both dataset. To make the comparison more meaningful, we also provide the FID values for the original images as a comparison. We calculated the FID of original images by randomly dividing the images into two groups. The score represents a level of best generation performance possible measured by FID. The FID quantities imply that our model generates more realistic images compared with the baseline models. More details regarding on how we calculate FID score can be found in Appendix A.4.

To quantitatively analyze the synthesized images with different noise injections, we computed the multi-scale structural similarity index measure (MS-SSIM) Wang et al. (2003). The MS-SSIM is an image quality assessment measurement that ranges from 0 to 1. It can measure the similarity of two input images. Higher MS-SSIM score indicates more similar results between two input images. Here, we randomly generated 10 images using our method and calculated the average MS-SSIM score between the original image and synthesized images to be 0.738. We also went through pairs of these 10 images and calculated the averaged MS-SSIM be-

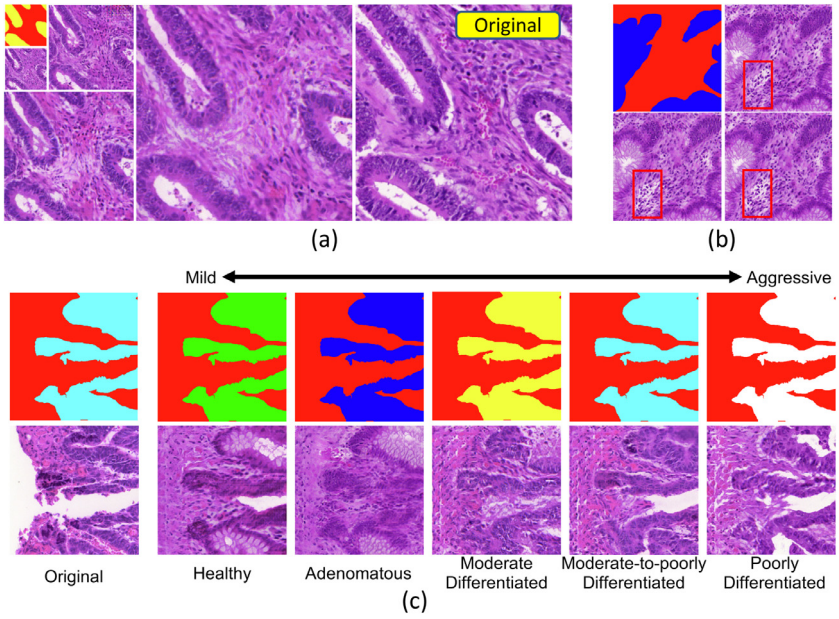

Fig. 4. (a) Generated coarse-to-fine results trained on the GlaS dataset. (b) Three generated images based on the same mask. Noise is injected during generation so that the model can synthesize images with variations. Clearer variations can been seen in the video clip in SI. (c) Image manipulation on synthesized images. Different gland types are observed when we change the label from healthy to poorly differentiated.

Table 2
GlaS challenge metrics for the total test set and subsets (A, B). * denotes methods that are operating on 512 × 512 scale.

Method	Object dice (A, B)	F1 score (A, B)	Hausdorff (A, B)
m-FCDenseNet*	0.748 (0.731, 0.792)	0.676 (0.662, 0.710)	123.39 (122.4, 125.9)
m-FCDenseNet + pyramid *	0.870 (0.894, 0.822)	0.860 (0.878, 0.776)	65.7 (54.1, 108.3)
Ours*	**0.874 (0.895, 0.825)**	**0.866 (0.890, 0.803)**	**61.85 (50.3, 100.4)**
FCN-8 Long et al. (2015)	0.781 (0.795, 0.767)	0.763 (0.783, 0.692)	124.2 (105.0, 147.3)
DeepLab Chen et al. (2017)	0.833 (0.859, 0.804)	0.813 (0.862, 0.764)	96.2 (65.7, 124.9)
Seg-Net Badrinarayanan et al. (2017)	0.838 (0.864, 0.807)	0.806 (0.858, 0.753)	92.6 (62.6, 118.5)
U-Net Ronneberger et al. (2015)	0.868 (0.884, 0.819)	0.841 (0.865, 0.768)	69.6 (55.6, 111)
DCAN Chen et al. (2016)	0.868 (0.897, 0.781)	0.863 (0.912, 0.716)	74.2 (45.4, 160.3)
Suggestive Annotation Yang et al. (2017)	NA (0.904, 0.858)	NA (0.921, 0.855)	NA (47.736, 96.976)
Xu et al. (2018)	NA (0.914, 0.859)	NA (0.930, 0.862)	NA (41.783, 97.390)
Graham et al. (2019)	**0.902 (0.919, 0.849)**	0.896 (0.920, 0.824)	54.7 (41.0, **95.7**)
Yan et al. (2020)	NA (0.914, 0.832)	NA **(0.927, 0.851)**	NA (**40.88**, 112.42)

tween them to be 0.768. Similarly, we calculated the results for pix2pixHD. The MS-SSIM score between synthesized images and the original image is 0.279, and the average score between synthesized images themselves is 0.317. As can be seen from these results, our method generates images with higher quality under noise injections than pix2pixHD.

4.3. Segmentation

We examine whether our proposed method can boost the performance in a fully-supervised segmentation task and reveal the contribution of each component to the performance through an ablation study.

4.3.1. Fully-supervised segmentation results

We first implemented the segmentation model as discussed in Section 3.2 in fully-supervised fashion. In supervised learning, we used the full training dataset, while using image synthesis to augment the training sets. Table 2 shows the segmentation performance of our model on the GlaS dataset compared with other studies. The performance shown in the table is based on the binary (stroma vs. epithelial glands) segmentation task, which is the same as in MICCAI Challenge 2015. We report the GlaS challenge metrics (Sirinukunwattana et al., 2017) including object-level F1 score, object-level dice coefficient, and object-level Hausdorff distance, and compare them with other main studies in literature. As shown in Table 2, we achieved comparable performance among all other studies, though our method does not surpass the state-of-the-art performance (Graham et al., 2019). We also present the segmentation results in Fig. 5 (row 1–3). We want to point out that instead of cropping the images to small patches, stitching back after the segment inference, and performing tedious post-processing as all other studies did, our method directly applies the segmentation model on the original 512 × 512 size. A direct comparison between our proposed method and m-FCDenseNet (backbone of S_n) reveals that our model is able to boost performance by a large margin when applied on high-res large-scale images directly. This difference is mainly due to the hierarchical structure of our model so that the effective receptive field of m-FCDenseNet is comparable with the state-of-the-art methods even with limited parameters. For the Prostate dataset, to make the results comparable with previous work (Li et al., 2017; Gertych et al., 2015; Ing et al., 2018), we used 5-fold cross validation with the standard metrics: mean Intersection Over Union (mIOU), Overall Pixel Accuracy (OPA) and Standard Mean Accuracy (SMA) to evaluate the performance of segmentation results. Assume we have segmentation results f, ground truth label l, and a pixel-wise confusion matrix $\boldsymbol{C}$, where $C_{i,j}$ is the number of pixels labeled as l_i and predicted as f_j. The mIOU is

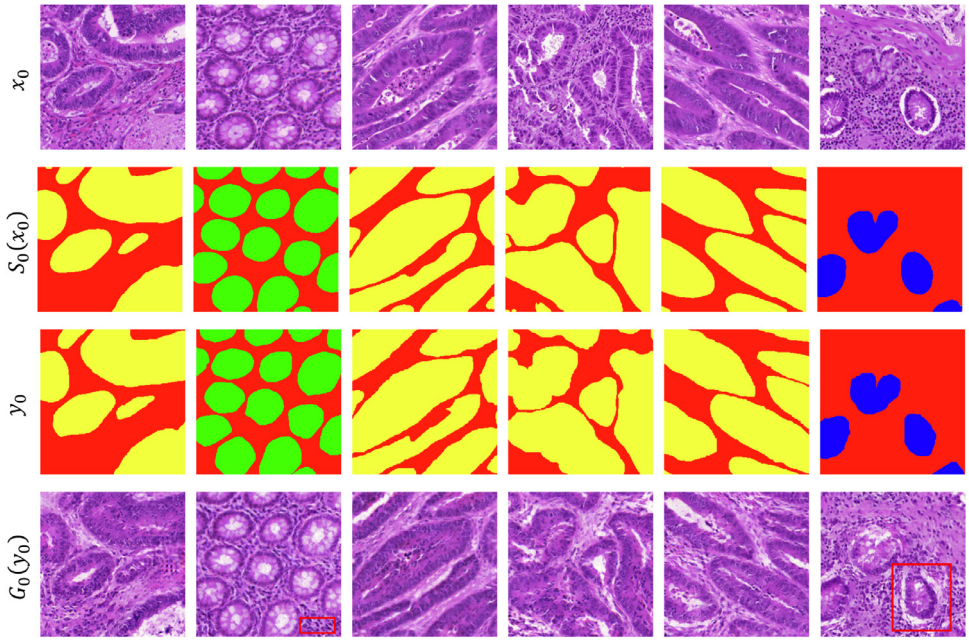

Fig. 5. Segmentation and generation results under fully-supervised scenario. x_0 are the original images; $S_0(x_0)$ are the semantic segmentation results by S_0; y_0 are the ground truth segmentation annotation; $G_0(y_0)$ are the synthetic images by G_0 conditioned on y_0.

Table 3
Model performance on segmenting prostate histological images as "Stroma" (BG), "Benign" (BN), "Low-Grade" (LG), and "High-Grade" (HG). * denotes methods that are operating on 512 × 512 scale.

Method	J_{BG}	J_{BN}	J_{LG}	J_{HG}	*mIOU*	*OPA*	SMA
m-FCDenseNet *	0.566	0.492	0.614	0.524	0.549	0.694	0.679
Ours *	**0.826**	**0.741**	**0.713**	**0.786**	**0.767**	**0.876**	**0.872**
Handcrafted Features Gertych et al. (2015)	0.595	0.352	0.495	N/A	0.481	N/A	N/A
Multi-Scale U-Net Li et al. (2017)	0.824	0.721	0.587	0.784	0.729	0.873	0.860
FCN-8s Ing et al. (2018)	N/A	N/A	N/A	N/A	0.759	0.873	N/A
Path R-CNN Li et al. (2018)	**0.831**	**0.839**	**0.715**	**0.797**	**0.796**	**0.894**	**0.888**

defined as the average of individual Jaccard coefficients, $\mathcal{J}_i$, for all classes l_i. The OPA is defined as the average of percent of pixels that are classified correctly for all classes l_i, $OPA = \frac{\sum_i C_{i,i}}{\sum_i \sum_j C_{i,j}}$. The standard mean accuracy is defined as $SMA = \frac{1}{N} \sum_i \frac{C_{ii}}{\sum_j C_{ij}}$. Table 3 shows the segmentation performance of our model on Prostate dataset. Similar to the observations in the GlaS dataset, our model achieves competitive results.

4.3.2. Ablation study

As we mentioned above, our proposed fully-supervised segmentation model is different from the traditional segmentation method in two ways. First, our method consists of a pyramid structure for generation and segmentation. Therefore, we can perform segmentation on large histopathology images without tiling the image and stitching it back together. Thanks to the pyramid structure, the final segmentation network S_0 has a larger receptive field that makes it able to consider both the large gland structures and finer nuclear details simultaneously. By skipping the tiling process, we are also able to avoid splitting the gland structure into different tiles and deteriorating the prediction accuracy. Second, our method leverages the synthetic images from G_n as augmented data to train S_n. It enlarges the training set size and is expected to improve the segmentation performance.

To determine how these two aspects affect the final performance, we conduct an ablation study. As shown in Table 2 of column "100%", the first three rows are all operating on 512 × 512 images. In first row *m-FCDenseNet*, we only have a single level S_0 applied to 512 × 512 image. Due to the limited receptive field, *m-FCDenseNet* alone has the worst segmentation performance. The *m-FCDenseNet+pyramid* model (row two) has the same pyramid structure as our model except that the synthetic images are not used to train S_n. Compared with *m-FCDenseNet* alone, the pyramid structure can effectively enlarge the receptive filed, leading to a performance improvement by a large margin. Conversely, comparing *m-FCDenseNet + pyramid* (row two) and our full model (row three), we observe the improvement is marginal, which indicates the synthetic images are not the key for performance boost in fully-supervised settings. We will further discuss this issue in Section 5.3.

4.4. SSL-segmentation

In this section, we examine whether our proposed method can boost the performance in a semi-supervised segmentation task and analyze how the pyramid structure and synthetic data augmentation contribute to the final performance.

4.4.1. SSL-segmentation results

We implemented SSL-segmentation and evaluated it by varying the amount of labeled data provided for training as commonly done in the literature (Chongxuan et al., 2017; Li et al., 2019). The labeled data used for training was randomly selected. In our experiments we used 20%, 40%, 60%, 80% labeled data and the rest as unlabeled data to train the model. Since there is no other public code implementation of the state-of-the-art methods that can

Table 4
mIOU for SSL-segmentation on GlaS dataset. All the results are generated under inductive learning unless specified in the parenthesis.

Task	Method	20%	40%	60%	80%
Binary	m-FCDenseNet	0.674	0.686	0.701	0.750
	m-FCDenseNet + pyramid	0.717	0.730	0.767	0.806
	Consistency-based	0.731	0.759	0.796	0.801
	Consistency-based (transductive)	0.737	0.777	0.811	0.821
	Ours	0.793	0.799	0.817	0.823
	Ours (transductive)	**0.817**	**0.824**	**0.845**	**0.830**
Multi-Category	m-FCDenseNet	0.203	0.254	0.262	0.282
	m-FCDenseNet + pyramid	0.216	0.259	0.289	0.325
	Consistency-based	0.228	0.267	0.310	0.331
	Consistency-based (transductive)	0.256	0.281	0.339	0.349
	Ours	0.242	0.287	0.301	0.336
	Ours (transductive)	**0.325**	**0.341**	**0.386**	**0.375**

Table 5
Object Dice for SSL-segmentation on GlaS dataset. All the results are generated under inductive learning unless specified in the parenthesis.

Task	Method	20%	40%	60%	80%	100%
Binary	m-FCDenseNet	0.643	0.661	0.675	0.681	0.748
	m-FCDenseNet + pyramid	0.651	0.655	0.789	0.818	0.870
	Ours	0.678	0.699	**0.833**	0.837	0.874
	Ours (transductive)	**0.703**	**0.725**	0.815	**0.840**	**0.886**

Table 6
F1 Score for SSL-segmentation on GlaS dataset. All the results are generated under inductive learning unless specified in the parenthesis.

Task	Method	20%	40%	60%	80%	100%
Binary	m-FCDenseNet	0.572	0.586	0.604	0.616	0.676
	m-FCDenseNet + pyramid	0.570	0.601	0.716	0.825	0.860
	Ours	0.596	0.593	**0.843**	0.839	0.866
	Ours (transductive)	**0.624**	**0.696**	0.776	**0.844**	**0.894**

be used directly for our problem, we implemented the consistency regularization-based method ourselves. Due to the time complexity of uncertainty measure on large images, we only worked on a simple version that incorporated segmentation loss and consistency loss. Specifically, the goal of our semi-supervised segmentation framework is to minimize the following combined objective function,

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_S(f(x_i;\theta)) + \lambda \sum_{i=1}^{N+M} \mathcal{L}_C(f(x_i;\theta',\xi'), f(x_i;\theta,\xi)), \tag{9}$$

where $\mathcal{L}_S$ denotes the supervised segmentation loss (i.e. , cross-entropy loss), and $\mathcal{L}_C$ represents the unsupervised consistency loss for measuring the consistency between the prediction of the teacher model and the student model for the same input x_i under different perturbations ξ. Here we add noise to the input to serve as a type of perturbation. For comparison, we also used m-FCDenseNet as the network backbone for this implementation. Furthermore, we conducted both inductive learning and transductive learning for our models, where in transductive learning the images in the test set were also treated as unlabeled data points for training. For the GlaS dataset, we performed two sets of SSL experiments: one for a binary segmentation task, where we only used a binary mask (stroma v.s epithelial gland) for training; and the other for a six-category segmentation task, where the mask not only contains the information of gland location but also the type of gland (healthy, adenomatous, moderately differentiated, moderate-to-poorly differentiated, or poorly differentiated). We report mean intersection over union (mIOU) as a metric to evaluate the segmentation performance for both binary and non-binary cases and demonstrate them in Table 4. In addition, we report Glas metrics for the binary case as shown inTables 5–7 since GlaS metrics are specifically designed for binary segmentation (glands vs. non-glands). We also performed a SSL-Segmentation test on the Prostate dataset. In order to reduce experimental time, instead of doing 5-fold cross-validation, we kept the 20% testing dataset fixed and report our evaluation on it. For each experiment discussed using the Prostate dataset, we ran it five times with different random seeds and report the mean and variance. The results are presented in Table 8. Note that all the results are generated under inductive learning unless specified in parentheses. As can be seen, our model outperforms the m-FCDenseNet and consistency-based method in all the cases of varying the amount of training data, demonstrating the effectiveness of our model. We also observe that our method improves the performance most with the low data regime as compared to the consistency-based method. Transductive learning outperforms inductive learning in most cases, which is expected since transductive learning incorporates the testing data as unlabeled data in training. Moreover, we found that the variance of model performance increased as we lowered the amount of training data, i.e. , the performance variance was larger when we only used 20% training data compared to the whole training set. This indicates the importance of selecting informative labeled data in the initial training stage.

4.4.2. Ablation study

To determine the effectiveness of the pyramid structure and synthetic data augmentation in the SSL setting, we present the binary segmentation results in column chart as shown in Fig. 6. In this figure, we only focus on inductive learning cases for fair comparison. Specifically, we calculate the performance gap between the single m-FCDenseNet baseline and our model and define two δ's as δ_1 to be *normalized performance improvement between* m-FCDenseNet *and* m-FCDenseNet + pyramid, and δ_2 to be *normalized*

Table 7
Hausdoff for SSL-segmentation on GlaS dataset. All the results are generated under inductive learning unless specified in the parenthesis.

Task	Method	20%	40%	60%	80%	100%
Binary	m-FCDenseNet	160.30	150.35	143.98	130.34	123.39
	m-FCDenseNet + pyramid	152.14	133.93	116.35	95.39	65.70
	Ours	139.87	122.66	115.28	105.44	**61.85**
	Ours (transductive)	**125.69**	**113.79**	**92.06**	**70.07**	62.21

Table 8
mIOU for SSL-segmentation on Prostate dataset. All the results are generated under inductive learning unless specified in the parenthesis.

Method	20%	40%	60%	80%
m-FCDenseNet	0.387 ± 0.037	0.420 ± 0.027	0.406 ± 0.025	0.492 ± 0.012
m-FCDenseNet + pyramid	0.437 ± 0.051	0.500 ± 0.039	0.581 ± 0.029	0.661 ± 0.017
Consistency-based	0.470 ± 0.076	0.499 ± 0.042	0.612 ± 0.070	0.658 ± 0.023
Consistency-based (transductive)	0.497 ± 0.056	0.531 ± 0.040	0.629 ± 0.037	0.667 ± 0.028
Ours	0.527 ± 0.063	0.573 ± 0.045	0.632 ± 0.030	0.694 ± 0.021
Ours (transductive)	0.577 ± 0.045	0.620 ± 0.028	0.664 ± 0.027	0.721 ± 0.013

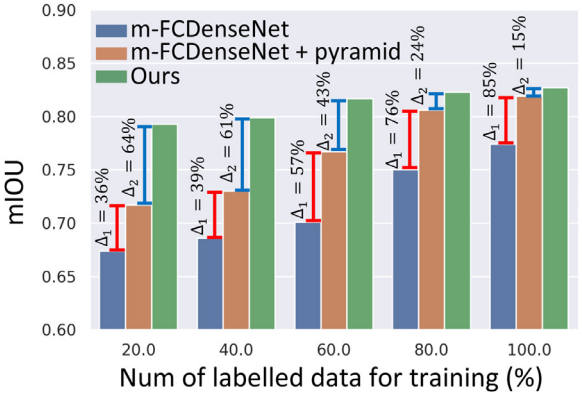

Fig. 6. Analysis of SSL-segmentation results on GlaS dataset. The experiments are done by using 512×512 images for the binary segmentation task.

performance improvement between m-FCDenseNet + pyramid *and* our method. Intuitively, δ_1 roughly characterizes the contributions from the pyramid structure and δ_2 roughly characterizes the contribution from synthetic data augmentation. As can be observed, δ_1 gradually increases as we increase the amount of labeled data, while δ_2 gradually decreases. This result indicates that in the low-labeled data scenario, synthetic data augmentation plays a more important role than the pyramid structure. As we have more labeled data, the pyramid structure becomes the key factor of performance improvement compared with m-FCDenseNet. Similar trends have been observed in other SSL-experiments (see Appendix A.3). We will further discuss the effectiveness of synthetic data augmentation in Section 5.3.

5. Discussion

5.1. Image generation

As shown both quantitatively (Table 1) and qualitatively (Fig. 3), our model generates more realistic images compared to the baseline models. The pix2pix method only leverages traditional conditional generative adversarial networks with a single Markovian discriminator (PatchGAN). It is more suitable for small-size images synthesis since the receptive field for both the generator and discriminator is limited. As a result, the generation of local patches is unaware of the global structure, potentially leading to spatial inconsistency. Pix2pixHD adopts a multi-scale generator and discriminator to capture the image features of different scales. As a result, it generates more realistic images with higher spatial consistency (e.g. gland structures are distinguishable from stroma). In contrast, our proposed model provides a pyramid structure, such that different scales focus on generating features of different levels. By conditioning on the generated images of the previous scale, our model is able to add finer details to the generated images while preserving the gland structure based on the semantic masks. As a result, our synthetic images are better in spatial consistency (compared with pix2pix model), and sharper with finer details (compared with pix2pixHD model).

5.2. Image scales for segmentation

Input image patch size is a key factor for segmentation performance. Therefore, to achieve good performance in histopathology image segmentation, researchers often have to tile the large-scale histopathology images into small patches and design a network structure with suitable receptive field. As a result of tiling, gland structures are often split into different parts, which deteriorates the segmentation accuracy. One of the merits of our proposed method for segmentation is that our model is not as sensitive to the input image scales compared to the single model with fixed receptive field. To demonstrate this, we evaluated our model from 64×64, up to 512×512 images on the GlaS dataset for the binary segmentation task. We compared it with m-FCDenseNet, which has the same architecture at a single scale S_n. Fig. 7 illustrates the results. In general, as we increase the image size, we expect the segmentation accuracy to increase because the resolution becomes higher. As can be seen, our model performs similar as m-FCDenseNet on small scale images. Once we input 512×512 images, the performance of our model increases while m-FCdenseNet suffers a sharp drop (see the red rectangle in Fig. 7). Since our model is relatively insensitive to the input image size, it is able to process large histopathology images with high magnification without breaking the gland structures into different tiles, which has been demonstrated to improve the segmentation accuracy in both fully-supervised and semi-supervised scenarios (see ablation study in Sections 4.3 and 4.4).

5.3. Effectiveness of synthetic data

As briefly discussed in Sections 4.3 and 4.4, we found that the synthetic data augmentation is not always helpful. In general, synthetic data augmentation is more effective in SSL, especially when the labeled images are extremely limited. These observations are

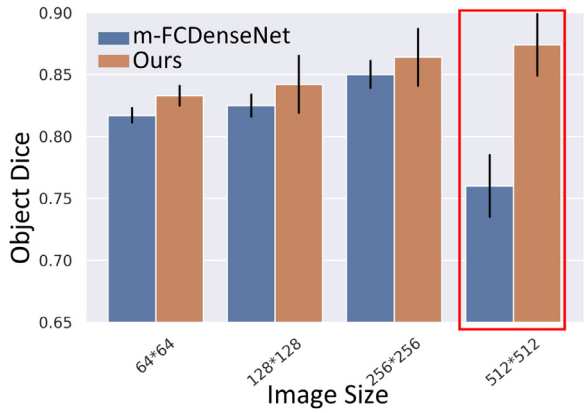

Fig. 7. Model performance with different input image scales. Our model is less sensitive to image scale compared to single-scale model such as m-FCDenseNet.

aligned with other studies (Goodfellow, 2016), where people find that generating a lot of additional samples by GAN and use them to provide a bigger dataset to train a classifier does not improve performance. The reason behind it is that it requires the generator to generalize better than the classifier, which is hard to achieve if you train both the generator and classifier on the same dataset. In this case, the generator gets no extra information compared to the classifier. It explains why our proposed model with synthetic data augmentation does not provide significant improvement compared to the *m-FCDesnseNet + pyramid* baseline in supervised settings. In fact, we argue that the normal data augmentation methods, such as flipping, rotation, color jitters etc. , are enough to provide strong regularization. Though the synthetic data augmentation provide extra regularization, the performance improvement is almost negligible (see Tables 2 and 4 in column "100%").

On the contrary, synthetic data augmentation works well in SSL as discussed in Section 4.4. Specifically, We generate pseudo masks through S_n for some unlabeled data and randomly choose these pairs as positive samples of D_n. This process introduces some bias to the target distribution of D_n, but it gives D_n a better chance to model the complete data distribution. In return, it enables G_n to explore a much larger image-mask manifold that includes both the labeled and unlabeled data information. In other words, G_n generated image-mask pairs are able to provide extra information gains compared with the labeled training sets. It also explains why we observe δ_2 is larger in the low-data scenario, while it is negligible when we use 100% labeled data for training. Intuitively, the larger information gains G_n can provider, the bigger improvement the synthetic data augmentation can contribute.

5.4. Limitations and future work

5.4.1. Better train-test data split

One limitation of our work is that the train-test data split used in our experiments is not a patient-wise validation. Unfortunately, we did not have patient-level information with which to perform a more rigorous patient-level stratification. This might result in a positive bias since a cancer can look similar in tiles within the same patient, especially in tiles that are spatially close to one another. Due to this reason, all performance numbers should thus be taken as relative and not absolute. However, as noted in Section 4.1, we argue that relative model comparisons in this work are fair as other works use similar split strategies.

5.4.2. Exploring meaningful latent representations

Besides the image generation results, we also provide image manipulation results by changing the gland labels in Fig. 4(c). It suggests that the generator G_n is not memorizing the training data itself but learning useful representations that are predictive for clinically relevant measurements. Nevertheless, in this study we do not provide any quantitative analysis on the learned representations. In future work, we plan to make the latent representations more explainable and associate them with clinically relevant measurements through mutual information maximization. We also plan to seek help from pathologists to provide clinically relevant measurements in future work.

5.4.3. Increasing memory efficiency

At each scale, our proposed model consists of three players, and because the algorithm has to maintain the weights of previous scales during training, the current model can occupy a lot of memory in GPU. Currently, when implementing the algorithm in a single Tesla V100S GPU, we are only able to generate 1024×1024 images using G_n and D_n alone. By incorporating S_n, we can only process image with sizes up to 512×512 at best. Therefore, potential future work is to increase the memory efficiency of the proposed method. It can be improved by two ways: first, we can take advantage of more memory-efficient network modules as backbone, such as EfficientNet etc. ; second, we can develop a random selection process to make the finer scale only focus on a sub-volume of images.

5.4.4. Improving segmentation results

Our model is complementary to the current state-of-the-art methods. For instance, the rotated convolution kernels used by Graham et al. (2019) can also be applied to S_n to increase performance. In the future, we would like to incorporate other ideas to improve the segmentation performance.

Additionally, we also found that there are generation artifacts in the synthetic images. For example, the synthetic image may miss a small part of a gland (see Fig. 5 column 2 red rectangular area), or the synthetic gland may not have a clear boundary as the real gland does (see Fig. 5 column 4 red rectangular area). These artifacts can potentially deteriorate the performance of segmentation, as we use the ground truth mask along with the synthetic images to train S_n. In the future, we would like to study the generation artifacts with the help of pathologist and improve the quality and diversity of synthetic images.

5.4.5. Investigating active learning

We performed multiple runs for SSL-training on the prostate dataset (see in Appendix A.3). For each run, we randomly selected a number of images as labeled data with the rest as unlabeled. We found that the variance of model performance increased as we decreased labelled training data, i.e. the performance variance was larger when we only used 20% training data compared to the whole training set. It indicates the importance of selected labeled data in the initial training stage. The results are not surprising and are related to active learning. In active learning, we have to develop a model to identify the most "important," "typical" images for expert to annotate. In this way, we can stabilize the performance in the low-labelled data regime. A potential future work could be extending our model for active learning.

6. Conclusion

In this work, we present a novel pyramid framework for synthesizing high-res histopathology images and use it to augment a dataset for a segmentation task in both supervised and semi-supervised scenarios. We provide detailed analysis on our synthetic images both qualitatively and quantitatively. We also demonstrate how the pyramid structure and synthetic data augmentation contribute to the final model performance differently. We conclude

that GANs can be effectively used to augment small pathology datasets to improve semantic segmentation in semi-supervised settings, which could potentially enhance downstream clinical analysis. We anticipate our findings can shed the light to the future researches on low-cost, high-res, large-scale histopathology image analysis.

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.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102251

CRediT authorship contribution statement

Wenyuan Li: Conceptualization, Methodology, Writing – original draft. **Jiayun Li:** Conceptualization, Methodology. **Jennifer Polson:** Writing – review & editing. **Zichen Wang:** Supervision, Conceptualization, Writing – review & editing. **William Speier:** Supervision, Conceptualization, Writing – review & editing. **Corey Arnold:** Supervision, Conceptualization, Writing – review & editing.

References

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: medical image translation using GANs. Comput. Med. Imaging Graph. 79, 101684.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MRimage segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 253–260.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., van de Kaa, C. H., Litjens, G., 2019. Automated Gleason grading of prostate biopsies using deep learning. arXiv preprint arXiv:1907.07980

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2020. Uncertainty aware temporal-ensembling model for semi-supervised ABUS mass segmentation. IEEE Trans. Med. Imaging 40 (1), 431–443.

Chen, H., Qi, X., Yu, L., Heng, P.-A., 2016. DCAN: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chongxuan, L., Xu, T., Zhu, J., Zhang, B., 2017. Triple generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 4088–4098.

Costa, P., Galdran, A., Meyer, M. I., Abràmoff, M. D., Niemeijer, M., Mendonça, A. M., Campilho, A., 2017. Towards adversarial retinal image synthesis. arXiv preprint arXiv:1701.08974

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi–supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.

Gertych, A., Ing, N., Ma, Z., Fuchs, T.J., Salman, S., Mohanty, S., Bhele, S., Velásquez–Vacca, A., Amin, M.B., Knudsen, B.S., 2015. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. Comput. Med. Imaging Graph. 46, 197–208.

Goodfellow, I., 2016. NIPS 2016 tutorial: generative adversarial networks. arXiv preprint arXiv:1701.00160

Graham, S., Epstein, D., Rajpoot, N., 2019. Rota-net: rotation equivariant network for simultaneous gland and lumen segmentation in colon histology images. In: European Congress on Digital Pathology. Springer, pp. 109–116.

Guibas, J. T., Virdi, T. S., Li, P. S., 2017. Synthetic medical images from dual generative adversarial networks. arXiv preprint arXiv:1709.01872

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein GANs. arXiv preprint arXiv:1704.00028

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.-S., Qin, J., 2020. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 562–571.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637.

Ing, N., Ma, Z., Li, J., Salemi, H., Arnold, C., Knudsen, B.S., Gertych, A., 2018. Semantic segmentation for prostate cancer grading by convolutional neural networks. In: Medical Imaging 2018: Digital Pathology, vol. 10581. International Society for Optics and Photonics, p. 105811B.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.

Lahiri, A., Ayush, K., Kumar Biswas, P., Mitra, P., 2017. Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: automated vessel segmentation in retinal fundus image as test case. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 42–48.

Lecouat, B., Foo, C.-S., Zenati, H., Chandrasekhar, V. R., 2018. Semi-supervised learning with GANs: revisiting manifold regularization. arXiv preprint arXiv:1805.08957

Li, J., Sarma, K.V., Ho, K.C., Gertych, A., Knudsen, B.S., Arnold, C.W., 2017. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. In: AMIA Annual Symposium Proceedings, vol. 2017. American Medical Informatics Association, p. 1140.

Li, W., Li, J., Sarma, K.V., Ho, K.C., Shen, S., Knudsen, B.S., Gertych, A., Arnold, C.W., 2018. Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images. IEEE Trans. Med. Imaging 38 (4), 945–954.

Li, W., Wang, Z., Yue, Y., Li, J., Speier, W., Zhou, M., Arnold, C. W., 2019. Semi-supervised learning using adversarial training with good and bad samples. arXiv preprint arXiv:1910.08540

Li, Y., Shen, L., 2018. cC-GAN: a robust transfer-learning framework for HEp-2 specimen image segmentation. IEEE Access 6, 14048–14058.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 523–534.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T., 2018. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 1038–1042.

Mahmood, F., Borders, D., Chen, R., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J., 2019. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE Trans. Med. Imaging 39 (11), 3257–3267.

Mok, T.C., Chung, A.C., 2018. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In: International MICCAI Brainlesion Workshop. Springer, pp. 70–80.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., Garnavi, R., 2019. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 282–290.

Senaras, C., Niazi, M.K.K., Sahiner, B., Pennell, M.P., Tozbikian, G., Lozanski, G., Gurcan, M.N., 2018. Optimized generation of high-resolution phantom images using cGAN: application to quantification of Ki67 breast cancer images. PloS One 13 (5).

Shaham, T.R., Dekel, T., Michaeli, T., 2019. Singan: learning a generative model from a single natural image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4570–4580.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: the glas challenge contest. Med. Image Anal. 35, 489–502.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nu-

clei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1196–1206.
Springenberg, J. T., 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390
Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807.
Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018b. pix2pixhd: high-resolution image synthesis and semantic manipulation with conditional GANs.
Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 542–551.
Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2. IEEE, pp. 1398–1402.
Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. 3D semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3646–3655.
Xu, X., Lu, Q., Yang, L., Hu, S., Chen, D., Hu, Y., Shi, Y., 2018. Quantization of fully convolutional networks for accurate biomedical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8300–8308.
Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SegAN: adversarial network with multi-scale l 1 loss for medical image segmentation. Neuroinformatics 16 (3–4), 383–392.
Yan, Z., Yang, X., Cheng, K.-T., 2020. Enabling a single deep learning model for accurate gland instance segmentation: a shape-aware adversarial learning framework. IEEE Trans. Med. Imaging 39 (6), 2176–2189.
Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: application to cross–modality liver segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 255–263.
Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 399–407.
Yi, X., Walia, E., Babyn, P., 2018. Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by Wasserstein distance for dermoscopy image classification. arXiv preprint arXiv:1804.03700
Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: areview. Med. Image Anal. 58, 101552.
You, C., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z., Cong, W., et al., 2019. Ct super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-circle). IEEE Trans. Med. Imaging 39 (1), 188–203.
Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.
Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 408–416.
Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A., 2019. Semi–supervised 3D abdominal multi-organ segmentation via deep multi-planar co–training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 121–140.