Style Consistent Image Generation for Nuclei Instance Segmentation

Xuan Gong[1], Shuyan Chen[1], Baochang Zhang[2], David Doermann[1]
[1]University at Buffalo [2]Beihang University
{xuangong, shuyanch, doermann}@buffalo.edu bczhang@buaa.edu.cn

Abstract

In medical image analysis, one limitation of the application of machine learning is the insufficient amount of data with detailed annotation, due primarily to high cost. Another impediment is the domain gap observed between images from different organs and different collections. The differences are even more challenging for the nuclei instance segmentation, where images have significant nuclei stain distribution variations and complex pleomorphisms (sizes and shapes). In this work, we generate style consistent histopathology images for nuclei instance segmentation. We set up a novel instance segmentation framework that integrates a generator and discriminator into the segmentation pipeline with adversarial training to generalize nuclei instances and texture patterns. A segmentation net detects and segments both real nuclei and synthetic nuclei and provides feedback so that the generator can synthesize images that can boost the segmentation performance. Experimental results on three public nuclei datasets indicate that our proposed method outperforms previous nuclei segmentation methods.

1. Introduction

In biomedical image analysis, instance detection and segmentation is an important task that assigns a semantic and object-level label at each pixel. A unique ID for each instance facilitates the further study of the nuclei spatial distribution to help understand the biological evolution of the cells concerning, for example, disease. In digital pathology, nuclear pleomorphism is required for tumor and cancer grading, and the spatial distribution of cancer nuclei is used as a prognostic for cancer prediction.

One challenge for deep learning is that it is data-hungry and often requires extensive and detailed annotation. For nuclei segmentation, both training and validation data typically require structured annotations, including bounding boxes and instance masks. These requirements increase the overall cost and time for data collection. For example, a training dataset consisting of 50 image patches (12M pixels) takes 120-230 hours of an expert pathologist's time [11]. A promising solution to the problem of the lack of well-labeled data is generating sample-label pairs. Generating synthetic data has been a hot topic recently and has been shown to significantly improve detection and classification performance [39, 9, 42]. The success of learning from synthetic images inspires us to study how to synthesize labeled nuclei images from existing annotations.

Nuclei images are intrinsically different from natural images and pose unique challenges for data generation. First, the distribution of sizes and shapes is quite diverse, and overlap and occlusion make the spatial variance more complicated. There are often large chromatic quality and density variations in hematoxylin and eosin (H&E) stained microscopy images. When datasets [21] are collected from multiple sites, differences of stain pigment and optical devices lead to a larger domain gap. Finally, widely used nuclei datasets [43, 21, 32] often combine the H&E stained images from different organ tissues, further increasing style variance, including stain quality and density. These variations of morphisms and style make automatic nuclei segmentation more difficult.

Some researchers have proposed learning a robust representation of nuclei images to incorporate the benefits of Generative Adversarial Networks (GANs) for synthesizing realistic pathology images from nuclei masks [11, 12, 30]. Current works usually synthesize nuclei using generative adversarial networks that demand a tremendous amount of data and elaborate manual work to generalize the appearance distribution (CycleGan [49]). However, there is still a gap between generating visually appealing images and improving the performance of downstream tasks like segmentation. For example, we may need extra focus on the boundary quality of generated samples for the segmentation task, but this is not a significant consideration for detection. In other words, there is no direct feedback from the downstream task network to the generator, which means the generator cannot be trained to target the specific downstream task.

We propose a style consistent adaptation module for the challenge of the diverse style of the nuclei images. Adver-

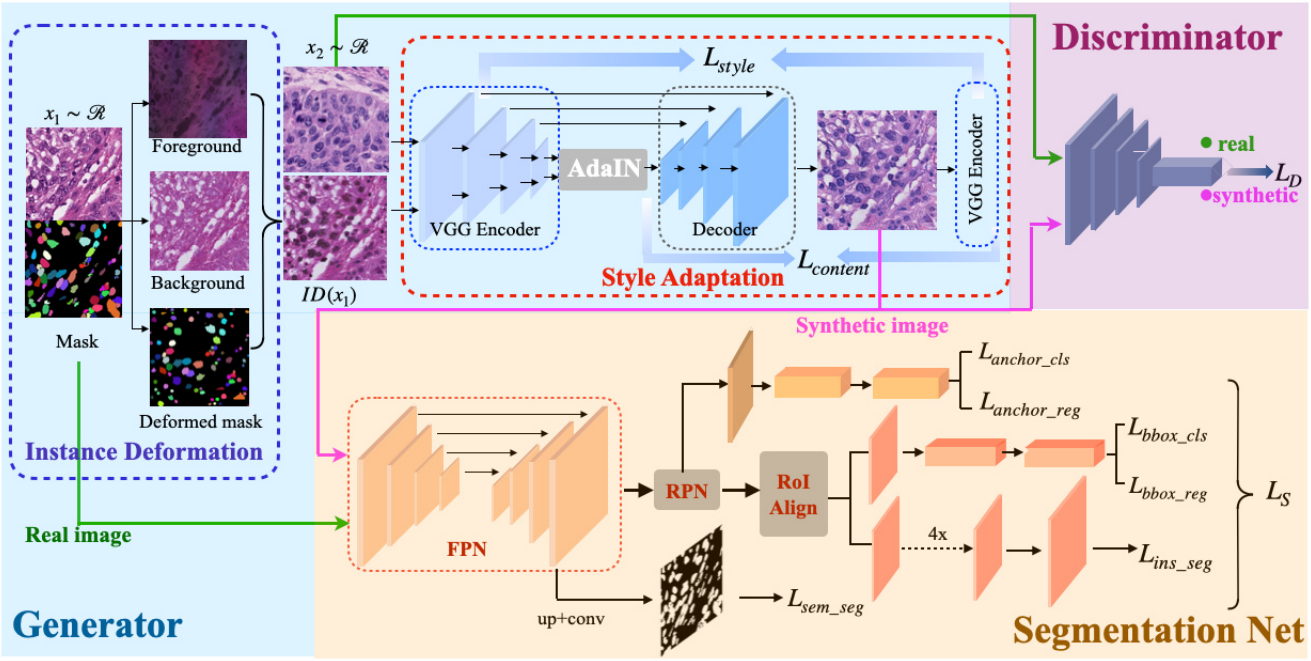

Figure 1. The framework of the proposed method. Our approach synthesizes images to help improve the nuclei instance segmentation. The segmentation net is integrated into the generator-discriminator loop.

sarial training is incorporated in the system to narrow the gap between visual-appealing-image generation and task-beneficial-image generation. In Fig. 1, we illustrate our proposed framework, which extends AdaIN [13] to a MaskR-CNN [8] based instance segmentation pipeline. We first combine nuclei deformation and adaptive instance normalization (AdaIN) in a generator to generalize the variability of nuclei pleomorphism and chromatic stain in the H&E images. The generator, discriminator, and instance segmentation pipelines are then integrated to synthesize nuclei images with less bias then real nuclei distributions. Previously proposed generative models only optimize for image realism and not for segmentation accuracy. To this end, we jointly optimize the generative model and segmentation network in an adversarial manner to provide a generalized representation of the nuclei and improve instance segmentation performance. Our contributions are summarized as follows:

1. We extend AdaIN to a generator, including instance deformation and style adaptation, to synthesize nuclei images with higher variances and realism, generating significantly better nuclear pleomorphisms and texture patterns.

2. We incorporate a generator into the instance segmentation pipeline so that the segmentation network can provide direct feedback to synthesize images that boost segmentation performance.

3. Experimental results on three nuclei histopathology datasets show that our approach leads to state-of-the-art performance compared with previous nuclei segmentation methods.

2. Related Work

2.1. Instance Segmentation

Deep learning-based instance segmentation is widely studied and can be categorized into proposal-free and proposal-based methods [7, 8, 45]. Proposal-free methods focus on morphology distribution and spatial relationships among all the objects in the images. Chen *et al*. [3] utilizes object boundaries to learn foreground probability maps and separate instances. Proposal-based methods are typically based on object detection. Mask R-CNN [8] uses a feature pyramid network (FPN) as the backbone to extract high-level features at multiple scales and feed them into a region proposal network (RPN) to generate regions of interest (ROIs). The ROIs are resized to fixed sizes and fed into a box branch and a mask branch to predict the class and mask separately. Following the fundamental framework of Mask R-CNN, panoptic segmentation [20, 19] proposes to unify semantic segmentation and instance segmentation.

For nuclei segmentation, panoptic segmentation is efficient and incorporates global semantic information. Cell R-CNN [4] jointly trains a semantic segmentation network and a Mask R-CNN with a shared backbone. Liu *et al*. [26] further designs a feature fusion module to incorporate global information during inference. Zhou *et al*. [48] explores in-

stance relationships and augments features from contextual information. Other improvements focus on the fine-grained segmentation around the boundary. [35] proposes a variance constrained cross-entropy loss that encourages the network to learn the spatial relationship between pixels in the same instance. [17] adds nuclei-boundary prediction as an intermediate step. We take advantage of both of these approaches in our segmentation pipeline, using internal layer supervision incorporating global information, and employing focal loss for fine-grained boundary determination.

2.2. Image-to-Image Translation

Facilitated by Generative Adversarial Networks (GANs) [6], conditional GAN [31, 15] dominates the task of image-to-image translation. Unpaired image-to-image translation methods [29, 49, 18] remove the requirement of paired-image supervision. Cycle-consistent Generative Adversarial Networks (CycleGan) [49, 14, 38], for example, enforces a bi-directional prediction between the source and target domain. AdaIN [13] designs simple yet efficient adaptive instance normalization to enable arbitrary style transfer with small scale training data. AdaIN has been used to generate person specific eyes from semantic mas [2].

Several works have proposed unsupervised approaches to synthesize histopathology images, due to the limited availability of labeled medical imaging data. Inspired by CycleGan, stainGan [37] eliminates stain color variation rather than performing for stain normalization. Re-staining Gan [47] is a CycleGan based method that transforms H&E stained images into immunohistochemistry (IHC) stained images. Mahmood utilizes CycleGan to transform content masks in the source domain and generate histopathology images as the target [30]. Similarly, Hu *et al*. [12] uses generative adversarial networks to learn a cell-level visual representation of histopathology images. They show that classification, segmentation, and detection can be carried out in an unsupervised manner with generative models. Hou *et al*. [11] fuses background and foreground with an instance mask and further refines the synthetic patch in a heterogeneous way.

2.3. Synthesis combined with Task

Some GAN models integrate auxiliary classifiers into the pipeline beyond the generator and discriminator's basic adversarial training [50]. Auxiliary Classifier GAN (ACGAN) [33] assesses the diversity of classes. Cycada [10] adds a task loss to the generative model for semantic segmentation. Ganin *et al*. [5] incorporates a domain classifier for the adversarial training. The methods intend to improve the generated image's realism for either a target class or a target domain, which is also known as domain adaptation [41].

Another approach is to utilize synthetic data to augment data for tasks such as detection and segmentation through adversarial training. A-Fast-RCNN [44] generates hard data augmentation transformations for the detector with adversarial training. Several works synthesize images for small-sized object detection [23], pedestrian detection [34, 46], and disease localization [28]. Liu *et al*. [27] applies adversarial domain adaptation to instance segmentation, but the aim is to adapt between two domains for unsupervised segmentation. Besides, image inpainting techniques [1, 40] are also used for synthesizing nuclei images [11].

For supervised instance segmentation, simply training the previously proposed generative models may yield minor improvements as they optimize image realism rather than task accuracy. To this end, we develop a framework that jointly optimizes the generative model and segmentation model such that the generated images improve the performance of instance segmentation.

3. Our Approach

Our method generates training images to directly improve instance segmentation performance.

3.1. Framework

The framework consists of a generator, an image-level discriminator, and an instance-level segmentation net (Fig. 1). The discriminator attempts to determine whether an image is real or synthetic. The instance segmentation net provides feedback to the generator about whether the generated images can improve segmentation performance.

Generator: Our generator includes a pre-synthesis mechanism and a refinement model based on adaptive instance normalization (AdaIN) [13]. The first step deforms nuclei instances as masks and merges a combination of foreground and background properties. The second step uses a generative model that enables style transfer from the initialized synthetic image to the real image based on training with a very limited number of images. The encoder is a fixed VGG-19 network to extract high-level features. The decoder learns to invert the output of adaptive instance normalization in the image space. We add a skip architecture [36] to preserve the content's high-frequency details to guarantee that the synthetic image is aligned with the deformed instance label. Our generator is very efficient at generating realistic nuclei images with only tens of training images. We will provide more details of the generator in Section 3.2.

Discriminator: As with a traditional discriminator in GANs, our discriminator attempts to classify an image is real or synthetic. We use the discriminator for super-resolution [22] to boost the synthetic images with realistic stain patterns at the image level. The discriminator learns to

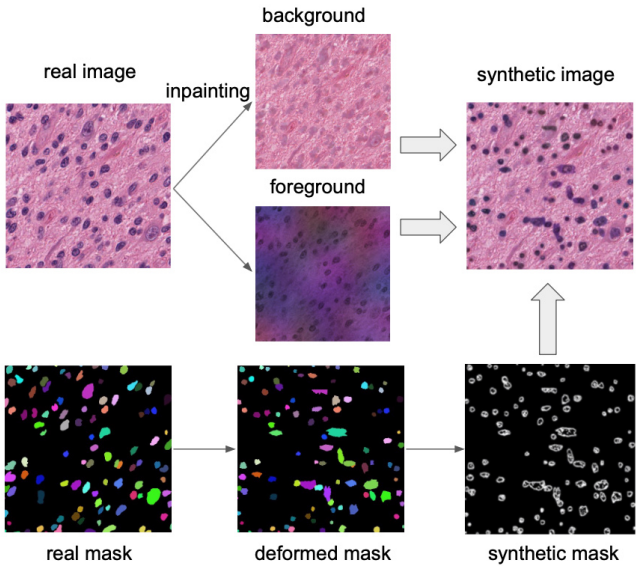

Figure 2. Illustration of instance deformation. The foreground and background are combined through deformed mask.

distinguish the synthetic nuclei images from real nuclei images, while the generator is optimized to synthesize realistic images to fool the discriminator.

Segmentation Net: This net is based on Mask R-CNN [8] for instance segmentation to evaluate instance realism. We add semantic supervision for panoptic image segmentation. The instance branch and semantic branch share the ResNet-50 feature pyramid network as the backbone. The network learns to detect and segment both real and synthetic nuclei instances. Like the discriminator, the result of segmentation is fed back to the generator and guides the synthesis at the instance level. The generator is optimized to synthesize instances that are difficult for the segmentation net to detect and segment. This mechanism of adversarial training makes the segmentation net more robust, thus improving the instance segmentation performance.

3.2. Image Synthesis

Histopathology image synthesis includes 1) a pre-synthesis step with deformation to generalize the spatial characteristics of nuclei instances such as size and shape and 2) a refine-synthesis step to adapt the synthesis images to the style of the real images. The pre-synthesis deforms the size and shape of the nuclei instances for an initial synthetic image. The refine-synthesis is to refine the initial synthetic image with the real image via adaptive instance normalization.

Deformation. As illustrated in Fig. 2, we utilize real histopathology images' spatial and texture characteristics to generate initial synthetic images. The deformed images result from a background/foreground fusion in Hematoxylin, Eosin, DAB (HED) color space [11]. Given a real image

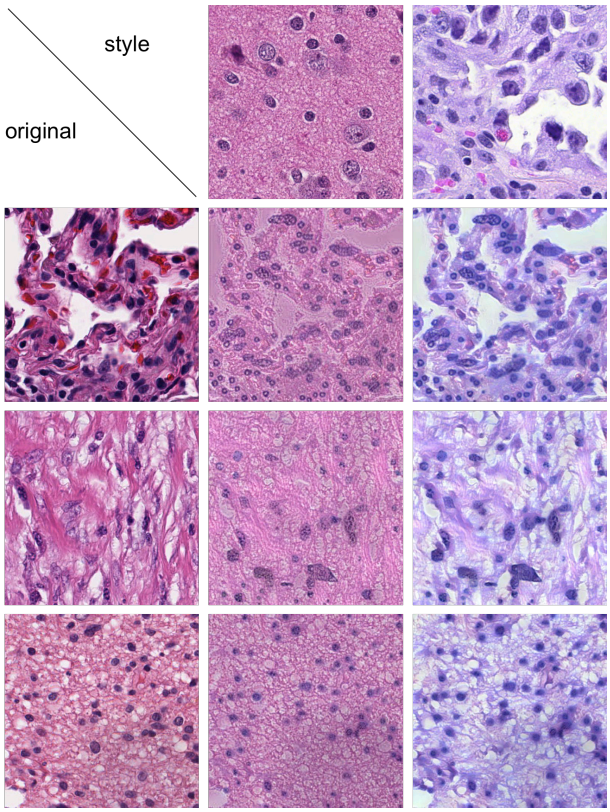

Figure 3. Illustrations of the synthetic images. The original image x_1 is used for instance deformation. The style image x_2 is to refine the initial synthetic image $ID(x_1)$.

x, we first create a background patch $B(x)$ by identifying and inpainting nuclei regions. Second, we simulate the nuclei's texture and intensity characteristics in the real images to create a foreground patch $N(x)$. We randomly resize and deform the contour of each instance and blur the instances. The foreground and background patches are augmented in the Hematoxylin and Eosin channels, respectively. Mask blur and patch augmentation are to make the fusion more realistic. Finally, we combine the augmented foreground and background patches based on the deformed instance mask $M(x)$ to produce the initial synthetic image $ID(x)$.

$$ID(x)_{i,j} = \begin{cases} B(x)_{i,j} & \text{if } M(x)_{i,j} = 0, \\ N(x)_{i,j} & \text{if } M(x)_{i,j} \neq 0, \end{cases} \tag{1}$$

where i,j is the pixel index, and $M_{i,j} = 0, ..., N$ is the background or instance index.

Synthesis. The fusion result based on the deformed instance may have some artifacts, so we use AdaIN [13] as the baseline of our generative model for style adaptation. The input includes one initial synthetic image $ID(x_1)$ as the content and one real image x_2 as the style. We denote the generated image as $\tilde{x} = G(x_1, x_2) = RS(ID(x_1), x_2)$. We use pre-trained layers before `relu4_1` of VGG-19 as

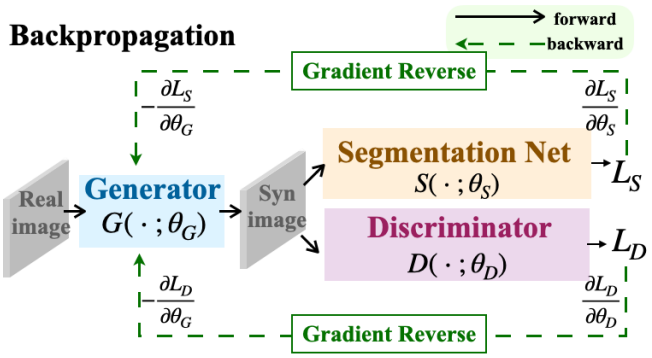

Figure 4. Overall backpropagation mechanism.

the encoder G_e to extract multi-scale features of content and style images. The content image's high-level features are adapted to the mean μ and variance σ of the content image to output t through adaptive instance normalization. We set up the decoder G_d with additional skip connections to ensure that the synthetic image $\tilde{x} = G_d(t)$ is consistent with $ID(x_1)$ in content and thus aligned with the corresponding instance label $\tilde{y} = M(x_1)$. The generation loss L_g is a combination of content loss $L_{content}$ and style loss L_{style}:

$$\begin{aligned} L_g &= L_{content} + \gamma L_{sytle} \\ &= ||G_e(G_d(t)) - t||^2 \\ &\quad + \gamma \sum_l ||\mu(\phi_l(\tilde{x})) - \mu(\phi_l(x_2))||^2 \\ &\quad + \gamma \sum_l ||\sigma(\phi_l(\tilde{x})) - \sigma(\phi_l(x_2))||^2, \end{aligned} \tag{2}$$

where $\phi_i (i = 1...4)$ denotes the output of layer `relu1_1`, `relu2_1`, `relu3_1`, and `relu4_1` in the VGG encoder, respectively. Fig. 3 illustrates the synthetic images $\tilde{x} = G(x_1, x_2)$. We will illustrate more synthetic images in the supplementary material.

3.3. Task guided Generative Models

The generator $G's$ objective is to generate images with deformed instances that are both realistic and help improve instance segmentation performance. Thus we design the discriminator D for realism evaluation and segmentation net S for instance segmentation. One of our main contributions is that the gradients derived from segmentation losses are backpropagated to the generator to boost the synthetic images, so they are useful for the instance segmentation. The generator's optimization is similar to the gradient reversal layer (GRL) [5] for adversarial training but at both image and instance levels. The overall backpropagation mechanism is illustrated in Fig. 4.

Algorithm 1 Training Strategy.

Input: generator G (including instance mask deformation M), discriminator D, segmentation net S, real histopathology images $\mathcal{R}$, mini-batchsize m.
Pretrain: Pretrain G and D using Eq. 4.
Pretrain S on $\mathcal{R}$ using Eq. 5.
for number of training iterations **do**
Sample mini-batch images $x_1, x_2 \in \mathcal{R}$
Forward G to get $\tilde{x} = G(x_1, x_2)$ and $\tilde{y} = M(x_1)$
Update G by ascending its stochastic gradient

$$\nabla_G \frac{1}{m} \sum_{(x_1,x_2)} [\log D(\tilde{x}) + L_s(S(\tilde{x}), \tilde{y}) + \lambda L_g(\tilde{x}, x_1, x_2))]$$

Update D by descending its stochastic gradient

$$\nabla_D \frac{1}{m} \sum_{(x_1,x_2)} [\log D(\tilde{x}) + \log(1 - D(x_2))]$$

Update S by descending its stochastic gradient

$$\nabla_S \frac{1}{m} \sum_{(x_1,x_2)} [L_s(S(x_1), y_1) + L_s(S(\tilde{x}), \tilde{y})]$$

end for

3.3.1 Adversarial Training at the Image Level

The generator G tries to synthesize realistic histopathology images that are indistinguishable with real nuclei images. Simultaneously, the discriminator D learns to classify real images as real and synthetic images as fake. We denote the distribution of the real domain and synthesis domain as $\mathcal{R}$ and $\mathcal{F}$, respectively. The adversarial loss at image level $L_{img\_adv}(G, D)$ can be written as:

$$\begin{aligned} &\max_G \min_D \mathop{\mathbb{E}}_{x\sim\mathcal{R}}[\log(1 - D(x))] + \mathop{\mathbb{E}}_{\tilde{x}\sim\mathcal{F}}[\log D(\tilde{x})] \\ =&\max_G \min_D \mathop{\mathbb{E}}_{x\sim\mathcal{R}}[\log(1 - D(x_2)) + \log D(G(x_1, x_2))]. \end{aligned} \tag{3}$$

Considering that the synthetic image should be consistent with the initial synthetic image $ID(x_1)$ in content and real image x_2 in style, the joint training of G and D at the global image-level becomes

$$L_{global}(G, D) = L_{img\_adv}(G, D) + \lambda L_g(G), \tag{4}$$

where λ is a balancing hyperparameter and is fixed during training.

3.3.2 Adversarial Training at Instance Level

Instance Segmentation. Based on Mask R-CNN, we use ResNet-50 as the feature pyramid network's backbone to extract features at multiple scales. The region proposal network focuses on instance detection and segmentation and these features are fed to a semantic segmentation branch

to learn semantic-level features. The semantic supervision directly applies to the backbone and facilitates the extraction of the most distinguishable features as the input to the region proposal network. Thus the additional semantic supervision contributes to the instance detection and segmentation. In addition, we replace the cross-entropy loss with a focal loss [24] setting $\lambda = 2$ for bounding box classification and instance segmentation. The aim is to give a larger penalty to the instances/pixels that are less distinguishable on dense images. Thus the loss for instance segmentation L_s is defined as:

$$\begin{aligned} L_s =& L_{anchor-cls(ce)} + L_{anchor-reg} + L_{bbox-cls(fl)} \\ &+ L_{bbox-reg} + L_{ins-seg(fl)} + L_{sem-seg(ce)}, \end{aligned} \tag{5}$$

where $L_{anchor-cls(ce)}$ and $L_{bbox-reg}$ are the losses for the classification (cross-entropy) and the anchor regression of the region proposal network (RPN), $L_{bbox-cls(fl)}$ and $L_{bbox-reg}$ are the classification loss (focal loss) and bounding box regression for the region of interest (ROI), $L_{ins-seg(fl)}$ is the focal loss for instance segmentation, and $L_{sem-seg(ce)}$ is the cross-entropy loss for semantic segmentation. We fuse the semantic features with the instance predictions during inference to output the final result, similar to the feature fusion mechanism in [26].

Adversarial Segmentation Loss: Given the generative model G for the generator and the net S for instance segmentation, the adversarial segmentation loss at the instance level $L_{ins\_adv}(G, S)$ can be written as:

$$\begin{aligned} &\max_G \min_S \mathop{\mathbb{E}}_{x\sim\mathcal{R}}[L_s(S(x), y)] + \mathop{\mathbb{E}}_{\tilde{x}\sim\mathcal{F}}[L_s(S(\tilde{x}), \tilde{y})] \\ =& \max_G \min_S \mathop{\mathbb{E}}_{x\sim\mathcal{R}}[L_s(S(x_1), y_1) + L_s(S(G(x_1, x_2)), M(x_1))], \end{aligned} \tag{6}$$

where y_1 is the pixel level instance annotation of the image x_1, and $\tilde{y} = M(x_1)$ is the instance annotation of the synthetic image $\tilde{x} = G(x_1, x_2)$. Segmentation net S is trained to minimize segmentation loss. The generator aims to synthesize images that can help to improve the performance of the instance segmentation. Thus G is optimized to maximize segmentation loss on synthetic images to generate images that the segmentation net has not seen and cannot predict well. The intuition is to generalize the nuclei instances and improve the robustness. Similar ideas have been proposed in DetectorGAN [28], which trains the generator adversarially to improve detection performance.

Alternatively, the generator could minimize the task-specific loss on synthetic images like ACGAN [33]. In this case, the auxiliary loss on synthetic images is minimized to improve the generator's realism. However, in our framework, the segmentation net should dominate since the goal is to improve instance segmentation performance. Synthetic instances may be biased away from the true data distribution and distract from the segmentation net's optimization [11]. Thus minimizing segmentation losses on synthetic images may not help and may even harm the segmentation performance on real images. Our experiments show that minimizing the task-specific losses on synthetic images like ACGAN decreases the segmentation performance on real images.

3.4. Overall Losses and Training

Our method should generate images that: (1) have instances that generalize the spatial characteristics of nuclei (pre-synthesis); (2) simulate the textures and styles of real histopathology images (refine-synthesis); (3) are indistinguishable from real images, both at the image (discriminator) and the instance (segmentation net) levels; and (4) contribute to the performance of instance segmentation. We have introduced all of these in our approach. The adversarial loss includes an image-level adversarial loss L_{img_adv} and an instance-level adversarial loss L_{ins_adv}.

$$\begin{aligned} &L_{adv} = L_{img\_adv}(G, D) + L_{ins\_adv}(G, S) \\ &= \min_D \mathop{\mathbb{E}}_{x_2\sim\mathcal{R}}[\log(1 - D(x_2))] + \min_S \mathop{\mathbb{E}}_{x_1\sim\mathcal{R}}[L_s(S(x_1), y_1)] \\ &\quad + \max_G \min_D \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{R}}[\log D(G(x_1, x_2))] \\ &\quad + \max_G \min_S \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{R}}[L_s(S(G(x_1, x_2), M(x_1)))]. \end{aligned} \tag{7}$$

The overall losses can be written as

$$L = L_{adv}(G, D, S) + \lambda L_g(G). \tag{8}$$

We pre-train the discriminator-generator pair (G, D) with global loss at the image level for faster convergence and pre-train the segmentation net S with real histopathology images. Note that we first pre-train (G, D) and S separately and then train them jointly. The optimization process is shown in Algorithm 1.

4. Experiments

4.1. Datasets

In our experiments, we use the following three datasets.

Cell17: The MICCAI 2017 Digital Pathology Challenge dataset [43] (Cell17) consists of H&E stained histology images. It contains 64 annotated images, and the training and testing sets contain eight images from four different diseases: glioblastoma multiforme (GBM), lower-grade glioma (LGG) tumors, head and neck squamous cell carcinoma (HNSCC), and non-small cell lung cancer (NSCLC). The image sizes are either 500×500 or 600×600 at $20\times$ or $40\times$ magnification.

TCGA: The MICCAI 2018 multi-organ segmentation challenge (MoNuSeg) used H&E stained tissue images captured at 40x magnification from the Cancer Genome Atlas (TCGA) archive. We refer to this as TCGA-kumar [21].

Method	AJI			Pixel-F1			Obj-F1		
	seen	unseen	all	seen	unseen	all	seen	unseen	all
Kumar *et al*. [32]	0.5154 ±0.0835	0.4989 ±0.0806	0.5083 ±0.0695	0.7301 ±0.0590	0.8051 ±0.1006	0.7623 ±0.0946	0.8226 ±0.0853	0.8322 ±0.0764	0.8267 ±0.0934
DIST [32]	0.5594 ±0.0598	0.5604 ±0.0663	0.5598 ±0.0781	0.7756 ±0.0489	0.8005 ±0.0538	0.7863 ±0.0550	-	-	-
Mask R-CNN [8]	0.5438 ±0.0649	0.5340 ±0.1283	0.5396 ±0.0929	0.7659 ±0.0481	0.7658 ±0.0608	0.7659 ±0.0517	0.6987 ±0.1344	0.6434 ±0.1908	0.6750 ±0.1566
Cell R-CNN [4]	0.5547 ±0.0567	0.5606 ±0.1100	0.5572 ±0.0800	0.7746 ±0.0446	0.7752 ±0.0577	0.7748 ±0.0485	0.7587 ±0.0969	0.7481 ±0.1488	0.7542 ±0.1166
Cell R-CNN v2 [26]	0.5758 ±0.0568	0.5999 ±0.1160	0.5861 ±0.0841	0.7841 ±0.0439	0.8078 ±0.0611	0.7943 ±0.0512	0.8014 ±0.0757	0.8023 ±0.1081	0.8017 ±0.0871
Cell R-CNN v3 [25]	0.5975 ±0.0568	0.6282 ±0.0924	0.6107 ±0.0726	0.7967 ±0.0453	0.8256 ±0.0520	0.8091 ±0.0487	0.8317 ±0.0694	0.8383 ±0.0598	0.8345 ±0.0631
Ours	0.6301 ±0.0696	0.6613 ±0.0633	**0.6346** ±0.0674	0.8159 ±0.0145	0.8305 ±0.0107	**0.8180** ±0.0131	0.8351 ±0.0304	0.8434 ±0.0276	**0.8379** ±0.0292

Table 1. Experimental comparisons on the TCGA dataset.

Method	Pixel-F1	Dice
Pix2Pix [16]	0.6208 ±0.1126	0.6351 ±0.0706
Mask R-CNN [8]	0.8004 ±0.0722	0.7070 ±0.0598
Cell R-CNN [4]	0.8216 ±0.0625	0.7088 ±0.0564
Liu *et al*. [26]	0.8645 ±0.0482	0.7506 ±0.0491
Ours	0.8622 ±0.0087	0.8216 ±0.0103

Table 2. Experimental comparisons on the Cell17 dataset.

The training sets consist of 30 annotated 1000×1000 patches and around 22,000 nuclei boundary annotations from the 30 slide images of different patients. These images show highly varying properties since they are from 18 different hospitals and seven different organs (breast, liver, kidney, prostate, bladder, colon, and stomach). For the test set, there are 14 images with additional 7000 nuclei boundary annotations.

TNBC: The Triple Negative Breast Cancer (TNBC) [32] dataset consists of 50 annotated 512×512 images at $40\times$ magnification with a total of 4022 annotated cells. The images are sampled from 11 patients at the Curie Institute. There are three to eight images for each patient. The image data includes low cellularity regions, which can be stromal areas or adipose tissue, and high cellularity areas consisting of invasive breast carcinoma cells.

4.2. Metrics

In our experiments, we evaluate detection, semantic segmentation, and instance segmentation. We use object-level F1-score as the detection metric:

$$F_1 = \frac{2TP}{FN + 2TP + FP}, \tag{9}$$

where TP, FN, and FP represent the number of true-positive (corrected detected objects), false-negative (ignored objects), and false-positive (detected objects without corresponding ground truth) detections with an IOU threshold of 0.5. The pixel-level F1-score is used as a semantic metric in segmentation. For the task of instance segmentation in histology images, the object-level Dice score [3] is one of the mainstream metrics to evaluate instance overlap and shape similarity, respectively. For the Cell17 dataset, we use the F1 score and Dice score to evaluate instance segmentation performance. For consistency with the current state-of-the-art TCGA and TNBC datasets, we employ the Aggregated Jaccard Index (AJI) [21] as the metric of instance segmentation. AJI computes an aggregated intersection cardinality numerator and an aggregated union cardinality denominator for all ground truth and segmented nuclei at the pixel level. We denote T_i as the binary mask of the ground truth nuclei instance i, P_j as the binary mask of the predicted nucleus instance j, and $J(i)$ as the index of predicted instances with the largest IOU with ground truth nucleus T_i:

$$J(i) = \underset{j \in \{j | j \neq J(i\prime)\}}{argmax} \frac{T_i \cap P_j}{T_i \cup P_j}, \tag{10}$$

where $i\prime = 1, ..., (i-1)$. In this way, each predicted instance can only be used once as $J(i)$. The AJI metric is written

Methods	AJI	Pixel-F1
Mask R-CNN [8]	0.5350 ±0.0993	0.7393 ±0.0977
Cell R-CNN [4]	0.5747 ±0.1061	0.7637 ±0.1080
Cell R-CNN V2 [26]	0.5986 ±0.0847	0.7793 ±0.0772
Cell R-CNN V3 [25]	0.6313 ±0.0750	0.8037 ±0.0557
Ours	**0.6316** ±0.0597	**0.8231** ±0.0137

Table 3. Experimental comparisons on TNBC dataset.

Training data			**Dice**	**Pixel-F1**
Real data	Syn Method	Adv training?		
✓	-	-	0.7823 ±0.0221	0.8501 ±0.0342
✓	CycleGan	✗	0.7213 ±0.0563	0.8033 ±0.0332
✓	Ours	✗	0.7487 ±0.0325	0.8156 ±0.0143
✓	CycleGan	✓	0.7842 ±0.0184	0.8471 ±0.0118
✓	Ours	✓	0.8216 ±0.0103	0.8622 ±0.0087

Table 4. Synthesis Comparison on Cell17 dataset.

Sem	Bbox FL	Mask FL	**Dice**	**Pixel-F1**
✗	✗	✗	0.7173 ±0.0276	0.8143 ±0.0398
✓	✗	✗	0.7594 ±0.0243	0.8447 ±0.0352
✗	✓	✗	0.7321 ±0.0274	0.8403 ±0.0373
✓	✓	✗	0.7667 ±0.0236	0.8476 ±0.0381
✓	✓	✓	0.7823 ±0.0221	0.8501 ±0.0342

Table 5. Ablation Study on Cell17 dataset. We report the results only trained with real data.

as

$$AJI = \frac{\sum_i |T_i \cap P_{J(i)}|}{\sum_i |T_i \cup P_{J(i)}| + \sum_{j \in \mathcal{S}} |P_j|}, \tag{11}$$

where $\mathcal{S} = \{j | j \neq J(i), \forall i\}$ is the set of false prediction instances without corresponding ground truth.

4.3. Experiments and Results

Cell17: In this experiment, we resize all the images to 512×512 and employ basic data augmentation, including horizontal and vertical flipping and rotations of 90^o, 180^o, and 270^o. We randomly crop the image into patches of size 256 and pre-train (G, D) pair using the same optimization settings as in AdaIN [13]. We pre-train the segmentation net S with Adam (lr=$1e-4$) with a batch size of 2 for 3000 iterations. The hyperparameter γ is 0.5 and λ is 10. The joint training optimizes with Adam (lr=$1e-5$) with a batch size of 1. The other Mask R-CNN based methods used for comparison [4, 26] use ResNet-101 as its backbone. In Tab. 2, we compare with existing semantic and instance segmentation models, including Pix2Pix [16], Mask R-CNN [8], Cell R-CNN [4] and the currents state-of-the-art [26].

TCGA: In this experiment, we randomly select one image from each organ in the training set for validation. During training, we crop the original 1000×1000 patches into four 512×512 patches. The training settings are the same as with Cell17. In addition to the basic data augmentation used in the experiments of Cell17, we add Gaussian blurring due to the high level of noise of this dataset. Note the TCGA test set images come from organ tissues such as breast, kidney, and lung, but lung examples are not seen in the training set. Thus we report the performances of images from seen organs and unseen organs separately in Tab. 1. The comparisons show that our method outperforms other works on both pixel level and instance level metrics.

TNBC: For our TNBC experiment, we use the same data split as the current state-of-the-art [26, 25]. We employ the same training settings as Cell17. The comparison in Tab. 3 shows that our method significantly outperforms others on pixel level F1 and achieves comparable performance on AJI.

4.4. Ablation Study

In Tab. 4, we show that simply training generative models does not yield satisfactory performance since they optimize for image realism rather than instance segmentation. We employ the same training setting for all the comparison methods training with synthetic data. And we use the same optimization setting with [26] for the setting where only real data is used for training. In addition, Tab. 5 compares the performances of segmentation net with and without semantic supervision, with cross-entropy and focal loss for bounding box classification, and with cross-entropy and focal loss for the instance mask. Note that we only train with real data for this comparison in Tab. 5. And we combine the semantic features with instance predictions during inference when semantic supervision is used for training.

5. Conclusion

In this work, we propose a style-consistent generation method for nuclei instance segmentation in histology images to deal with insufficient data. We generalize the spatial characteristics and stain patterns through instance deformation and style adaptation. Furthermore, we integrate the generator into the segmentation pipeline and optimizer the generator with adversarial training on the synthetic images. Experimental results show that the synthetic images help to improve the instance segmentation performance.

References

[1] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[2] Marcel Bühler, Seonwook Park, Shalini De Mello, Xucong Zhang, and Otmar Hilliges. Content-consistent generation of realistic eyes with style. *arXiv preprint arXiv:1911.03346*, 2019.

[3] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. DCAN: deep contour-aware networks for accurate gland segmentation. *CVPR*, 2016.

[4] D.Zhang, Y.Song, D.Liu, H.Jia, S.Liu, Y.Xia, H.Huang, and W.Cai. Panoptic segmentation with an end-to-end cell r-cnn for pathology image analysis. *MICCAI*, page 237–244, 2018.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, pages 1–19, 2020.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. 2017.

[9] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

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

[12] Bo Hu, Ye Tang, I Eric, Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. *IEEE journal of biomedical and health informatics*, 23(3):1316–1328, 2018.

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[14] Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. *COLING*, 2020.

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

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with con- ditional adversarial networks. *CVPR*, page 5967–5976, 2017.

[17] Qingbo Kang, Qicheng Lao, and Thomas Fevens. Nuclei segmentation in histopathological images using two-stage learning. *MICCAI*, page 703–711, 2019.

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[19] A. Kirillov, R. Girshick, K. He, and P. Dolla . Panoptic feature pyramid networks. *CVPR*, page 6399–6408, 2019.

[20] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dolla . Panoptic segmentation. *CVPR*, pages 9404–9413, 2019.

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

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[23] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1222–1230, 2017.

[24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *ICCV*, pages 2980–2988, 2017.

[25] Dongnan Liu, Donghao Zhang, Yang Song, Heng Huang, and Weidong Cai. Cell r-cnn v3: A novel panoptic paradigm for instance segmentation in biomedical images. 2020.

[26] D. Liu, D. Zhang, Y. Song, C. Zhang, F. Zhang, L. ODonnell, and W. Cai. Nuclei segmentation via a deep panoptic model with semantic feature fusion. *IJCAI*, pages 861–868, 2019.

[27] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4243–4252, 2020.

[28] Lanlan Liu, Michael Muelly, Jia Deng, Tomas Pfister, and Li-Jia Li. Generative modeling for small-data object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6073–6081, 2019.

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[30] Faisal Mahmood, Daniel Borders, Richard Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging*, 2019.

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[32] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2):448–459, 2018.

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[34] Xi Ouyang, Yu Cheng, Yifan Jiang, Chun-Liang Li, and Pan Zhou. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. *arXiv preprint arXiv:1804.02047*, 2018.

[35] Hui Qu, Zhennan Yan, Gregory M. Riedlinger, Subhajyoti De, and Dimitris N. Metaxas1. Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss. *MICCAI*, page 378–386, 2019.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[37] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 953–956. IEEE, 2019.

[38] Liyue Shen, Wentao Zhu, et al. Multi-domain image completion for random missing input data. *arXiv preprint arXiv:2007.05534*, 2020.

[39] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[40] Liangchen Song, Bo Du, Lefei Zhang, Liangpei Zhang, Jia Wu, and Xuelong Li. Nonlocal patch based t-svd for image inpainting: Algorithm and error analysis. In *AAAI Conference on Artificial Intelligence*, 2018.

[41] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102:107173, 2020.

[42] Liangchen Song, Yonghao Xu, Lefei Zhang, Bo Du, Qian Zhang, and Xinggang Wang. Learning from synthetic images via active pseudo-labeling. *IEEE Transactions on Image Processing*, 2020.

[43] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, Rajarsi Gupta, Jin Tae Kwak, Nasir Rajpoot, Joel Saltz, and Keyvan Farahani. Methods for segmentation and classification of digital microscopy tissue images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2018.

[44] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2606–2615, 2017.

[45] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1570–1578, 2020.

[46] Jialian Wu, Chunluan Zhou, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Temporal-context enhanced detection of heavily occluded pedestrians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13430–13439, 2020.

[47] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: a promising solution for whole slide image analysis. *arXiv preprint arXiv:1901.04059*, 2019.

[48] Yanning Zhou, Qi Dou3 Hao Chen, Jiaqi Xu1, and Pheng-Ann Heng. Irnet: Instance relation network for overlapping cervical cell segmentation. *MICCAI*, page 640–648, 2019.

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

[50] Wentao Zhu, Xiang Xiang, Trac D Tran, Gregory D Hager, and Xiaohui Xie. Adversarial deep structured nets for mass segmentation from mammograms. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 847–850. IEEE, 2018.