Robust Histopathology Image Analysis: to Label or to Synthesize?

Le Hou[1], Ayush Agarwal[1,2], Dimitris Samaras[1], Tahsin M. Kurc[1], Rajarsi R. Gupta[1], Joel H. Saltz[1]
[1]Stony Brook University [2]Stanford University, California
{lehhou,samaras}@cs.stonybrook.edu ayush94582@gmail.com
{tahsin.kurc,joel.saltz}@stonybrook.edu rajarsi.gupta@stonybrookmedicine.edu

Abstract

Detection, segmentation and classification of nuclei are fundamental analysis operations in digital pathology. Existing state-of-the-art approaches demand extensive amount of supervised training data from pathologists and may still perform poorly in images from unseen tissue types. We propose an unsupervised approach for histopathology image segmentation that synthesizes heterogeneous sets of training image patches, of every tissue type. Although our synthetic patches are not always of high quality, we harness the motley crew of generated samples through a generally applicable importance sampling method. This proposed approach, for the first time, re-weighs the training loss over synthetic data so that the ideal (unbiased) generalization loss over the true data distribution is minimized. This enables us to use a random polygon generator to synthesize approximate cellular structures (i.e., nuclear masks) for which no real examples are given in many tissue types, and hence, GAN-based methods are not suited. In addition, we propose a hybrid synthesis pipeline that utilizes textures in real histopathology patches and GAN models, to tackle heterogeneity in tissue textures. Compared with existing state-of-the-art supervised models, our approach generalizes significantly better on cancer types without training data. Even in cancer types with training data, our approach achieves the same performance without supervision cost. We release code and segmentation results[1] on over 5000 Whole Slide Images (WSI) in The Cancer Genome Atlas (TCGA) repository, a dataset that would be orders of magnitude larger than what is available today.

1. Introduction

Existing state-of-the-art supervised image analysis methods [11, 22, 13, 48, 3, 62, 59, 61, 9, 66, 64, 24, 40] largely rely on the availability of large annotated training datasets which requires the involvement of domain experts. This is a time-consuming and expensive process. Moreover, for methods that generalize on various input types, supervised data must be collected for every input type. For example, labeled satellite images from regions such as north Europe and south Africa are all needed to train a robust satellite image analysis method [65, 49]. In pathology image analysis, to achieve optimal performance, the data annotation phase often must be repeated for different tissue types such as different cancer sites, fat tissue, necrotic regions, blood vessels, and glands, because of tissue heterogeneity as well as variations in tissue preparation and image acquisition. The detection, segmentation, and classification of nuclei are core analysis steps in virtually all pathology imaging studies [11, 22, 13, 48, 3, 62, 59, 61, 9, 66, 64, 40, 23, 2, 29] and precision medicine [17, 12]. It is the first step in extracting interpretable features that provide valuable diagnostic and prognostic cancer indicators [14, 15, 1, 43, 20]. Manual generation of nucleus segmentation ground truth data takes a long time. In our experience, a training dataset consisting of 50 image patches (12M pixels) takes 120-230 hours of an expert pathologist's time. This training dataset is extremely small compared with the volume of data in a large study (*e.g.* 10k whole slide images, 50T pixels). This is a major impediment to robust nucleus segmentation.

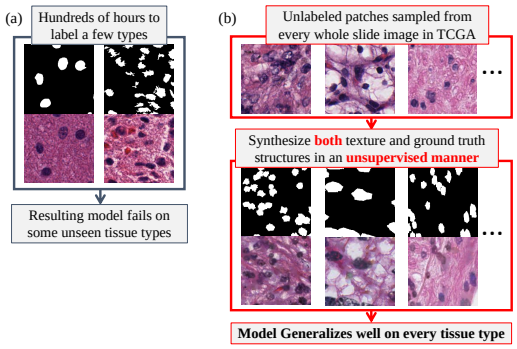


Figure 1. (a). Standard learning methods learn and perform well only with tissue types for which ground truth training data exists. (b). We propose to synthesize both image texture and ground truth structures for training a supervised model, even when no real ground truth structures are given. As a result, our model generalizes well on unseen tissue types.

[1] www3.cs.stonybrook.edu/~cvl/nuclei_seg.html

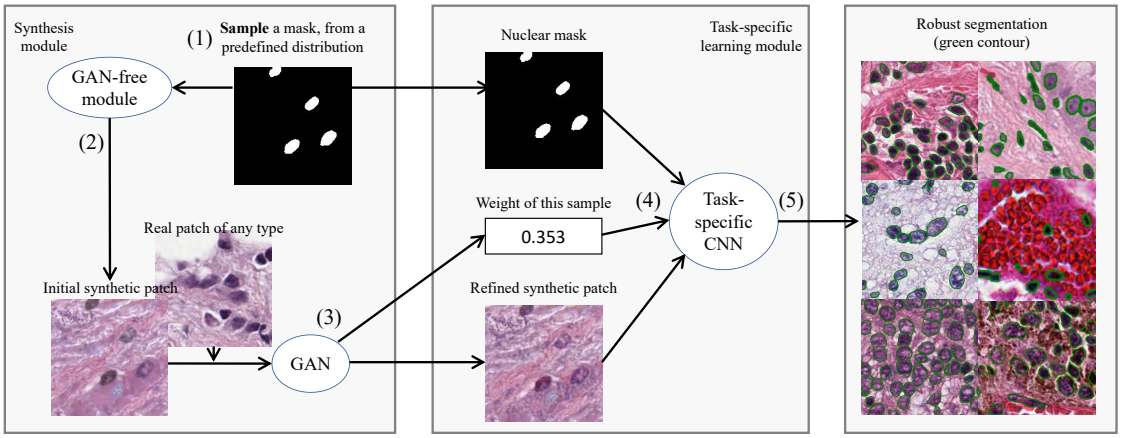


Figure 2. Overview of our pipeline: we use a GAN-free module to synthesize (sample) an initial synthetic pathology image patch with its nuclear mask. We then refine the initial synthetic patch using a GAN and compute its sample weight. We finally train a task-specific (*e.g.* segmentation, classification, *etc.*) CNN on this sampled instance. If a sampled ground truth structure does not produce a realistic synthetic example, the impact of this instance on the training loss is down-weighted.

One approach to address this problem is training data synthesis [26, 16, 51]. All existing training data synthesis approaches assume that the distribution of synthetic data is the same as the distribution of real data. However, this is often not the case, especially for synthesis of histopathology images with cellular structure (*e.g.* nuclear masks), since no real examples of nuclear masks are given for most cancer types. We propose an importance sampling based approach that minimizes the ideal (unbiased) generalization loss over the distribution of real data, even when given a biased distribution (of synthetic data). This allows us to enumerate possible cellar structures for training data synthesis. Our pipeline (see Fig. 2):

1. Samples a nucleus segmentation mask from a predefined, approximate ground truth generator;
2. Constructs an initial synthetic patch utilizing real textures (Fig. 3) of the input tissue type;
3. Uses a GAN model to make the initial synthetic patch more realistic;
4. Computes an importance weight of this synthetic example, from the discriminator's output simply using Bayes' theorem; and
5. Trains a task-specific (*e.g.* segmentation) CNN using the synthetic patch, mask and importance weight.

In other words, we enumerate possible ground truth structures during generation of synthetic training patches. If a resulting patch is not realistic, we decrease its impact in the training loss. Similarly, if a resulting patch is not only very realistic, but also rarely synthesized, then we increase its impact in the training loss.

To summarize, our contributions are: **(1)** Synthesizing perfectly realistic training patches with masks is almost impossible when we are not given any real examples of nuclear masks. We propose an importance sampling based method that reweighs the losses of approximately generated examples, for training a task-specific (*e.g.* nucleus segmentation) network, minimizing the ideal (unbiased) generalization loss over the real data distribution. **(2)** We show how to compute importance weights from the outputs of the GAN discriminator by simply using the Bayes' theorem, without any computational overhead. **(3)** We propose a hybrid synthesis pipeline that utilizes textures in real histopathology patches for synthesis of any tissue patches. **(4)** The proposed method is robust to tissue heterogeneity. When there are no supervised datasets for a test cancer type, our nucleus segmentation CNN significantly outperforms supervised methods in across-cancer generalization. Even for the few tissue types for which supervised data exist, our method matches the performance of supervised methods. **(5)** We release nucleus segmentation results on over 5000 Whole Slide Images (WSI) of 13 major cancer types in The Cancer Genome Atlas (TCGA) repository. These results are at least four orders of magnitude larger than currently available human annotated datasets. We believe that this large-scale dataset, even though not as accurately annotated, is a useful feature for future pathology image analysis research.

2. Related Work

Detection and segmentation of nuclei is a fundamental analytical step in virtually all pathology imaging studies [11, 22, 13, 48, 3, 62, 59, 61, 9, 66, 64, 40, 23, 2, 29] and precision medicine [17, 12]. Recent works in image analysis have proposed crowd-sourcing or high-level, less accurate annotations, such as scribbles, to generate large training datasets manually [34, 57, 64]. Work by Zhou *et al.* [68] segments nuclei inside a tissue image and redistributes the segmented nuclei inside the image. The segmentation masks of the redistributed nuclei are

assumed to be the predicted segmentation masks. This work requires segmentation masks and does not generate new textures and shapes. Generative Adversarial Networks (GANs) [44] have been proposed for generation of realistic images [16, 6, 4, 51, 8, 67, 42, 25, 46, 38]. For example, an image-to-image translation GAN [26, 16] synthesizes eye fundus images. However, it requires an accurate supervised segmentation network to segment eye vessels out, as part of the synthesis pipeline. The S+U learning framework [51] refines initially synthesized images via a GAN to increase their realism. This method achieves state-of-the-art results in eye gaze and hand pose estimation tasks. Recently, a GAN based approach [37] is able to synthesize realistic pathology images with nuclear masks. It is limited to cancer types with ground truth masks, since it requires real mask examples. GANs are also used to synthesize images of various styles of the same content. Cycle-GAN *etc*. [35, 69] transfers content of images to target styles without training with paired images. The universal style transfer approach [32, 54] solves this problem by providing a reference style to the generator network. However, to apply any of the GAN models for synthesizing image and masks, examples of both real images and masks are required.

3. Importance Sampling for Loss Estimation

In this section we show how to minimize the *ideal (unbiased)* task-specific (*e.g*. segmentation, classification, *etc*.) generalization loss over the distribution of real data, given an approximate sampling distribution (of synthetic data). We define a random variable X representing an image/patch, with its ground truth T, and the probability density function of real images as $\mathrm{p}(\langle X,T\rangle)$. In practice, X and T are discrete. The task-specific generalization loss $L_R(\theta_R)$ with model parameters θ_R is:

$$L_R(\theta_R) = \sum_{X,T} \mathrm{f}_{\theta_R}(\langle X,T\rangle)\mathrm{p}(\langle X,T\rangle), \tag{1}$$

where $\mathrm{f}_\theta(\cdot)$ is the loss function such as the conventional segmentation loss [36, 41]. To minimize the generalization loss defined by Eq. 1, we sample one example $\langle X,T\rangle$ from the distribution defined by $\mathrm{p}(\langle X,T\rangle)$, then minimize the loss $\mathrm{f}_\theta(\langle X,T\rangle)$. If there are infinite real samples, the empirical loss converges exactly to Eq. 1. In this work, we synthesize training examples $\langle X,T\rangle$. We define the probability density function of synthetic images as $\mathrm{g}(\langle X,T\rangle)$. Ideally $\mathrm{p}(\langle X,T\rangle)$ is equivalent to $\mathrm{g}(\langle X,T\rangle)$. However, for synthesizing unbiased examples *and corresponding "ground truth" nuclear masks*, an unbiased modeling of nuclear masks is needed – existing training image synthesis methods [51] heavily depend on unbiased ground truth image structure modeling, such as size of eyeballs, color of iris. This is almost impossible for histopathology images because of the paucity of annotated data and the cellular structure heterogeneity across tissue types.

To estimate the ideal (unbiased) generalization loss with $\mathrm{g}(\langle X,T\rangle)$, we formulate the task-specific loss as follows:

$$L_R(\theta_R) = \sum_{X,T} \mathrm{f}_\theta(\langle X,T\rangle)\frac{\mathrm{p}(\langle X,T\rangle)}{\mathrm{g}(\langle X,T\rangle)}\mathrm{g}(\langle X,T\rangle). \tag{2}$$

Instead of sampling $\langle X,T\rangle$ from the real pdf $\mathrm{p}(\langle X,T\rangle)$, we can now sample $\langle X,T\rangle$ from the synthetic pdf $\mathrm{g}(\langle X,T\rangle)$ and minimize a new loss function $\mathrm{f}'(\langle X,T\rangle) = \mathrm{f}_\theta(\langle X,T\rangle)\mathrm{p}(\langle X,T\rangle)/\mathrm{g}(\langle X,T\rangle)$. This is the standard importance sampling approach [7]: when sampling from $\mathrm{p}(\langle X,T\rangle)$ is expensive, we sample from $\mathrm{g}(\langle X,T\rangle)$ then re-weight each sample by multiplying its loss with weight $\mathrm{p}(\langle X,T\rangle)/\mathrm{g}(\langle X,T\rangle)$. Note that for the resulting generalization loss estimation to be unbiased, for all $\langle X,T\rangle$ with $\mathrm{p}(\langle X,T\rangle) > 0$, it is required that also $\mathrm{g}(\langle X,T\rangle) > 0$.

Given an image X, the underlying ground truth T is fixed. Thus, we can simply drop T in PDFs:

$$\frac{\mathrm{p}(\langle X,T\rangle)}{\mathrm{g}(\langle X,T\rangle)} = \frac{\mathrm{p}(X)}{\mathrm{g}(X)}. \tag{3}$$

The right hand side of Eq. 3 can be derived from the output of a GAN discriminator. A discriminator trained with cross-entropy (log-likelihood) loss estimates the probability that X is sampled from the real distribution instead of the synthetic distribution: $\Pr(X \sim \mathrm{p}|X)$. The discriminator is trained with real and synthetic examples. Denote a constant c as the ratio between the numbers of synthetic input samples and real input samples: $c = \Pr(X \sim \mathrm{g})/\Pr(X \sim \mathrm{p})$. Thus $\mathrm{p}(X) = \Pr(X|X \sim \mathrm{p})$, $\mathrm{g}(X) = \Pr(X|X \sim \mathrm{g})$. Using Bayes' theorem, we have:

$$\begin{aligned}\Pr(X \sim \mathrm{p}|X) &= \frac{\Pr(X|X \sim \mathrm{p})}{\Pr(X|X \sim \mathrm{p}) + \Pr(X|X \sim \mathrm{g})\,c} \\ &= \frac{\mathrm{p}(X)}{\mathrm{p}(X) + \mathrm{g}(X)c}.\end{aligned} \tag{4}$$

Rearranging Eq. 4 gives us the importance weight formulated by the discriminator's output $\Pr(X \sim \mathrm{p}|X)$:

$$\frac{\mathrm{p}(X)}{\mathrm{g}(X)} = c \cdot \frac{\Pr(X \sim \mathrm{p}|X)}{1 - \Pr(X \sim \mathrm{p}|X)}. \tag{5}$$

If a synthetic patch is unrealistic ($\Pr(X \sim \mathrm{p}|X) \ll 0.5$), it will be down-weighted (contribute less to the loss). If a synthetic patch is realistic and rarely generated, it will be up-weighted (contribute more to the loss). We show the visualization of importance weights in Fig. 7.

Optimality of unbiased loss minimization: Since we learn $\Pr(X \sim \mathrm{p}|X)$ via training the discriminator on the unbiased dataset (*i.e*. unlimited samples of $X \sim \mathrm{p}$ and $X \sim \mathrm{g}$), we can easily show that this yield unbiased generalization loss minimization: The unbiased generalization loss over the distribution of real data defined by Eq. 1 is

equivalent to Eq. 2. Since we can sample from the synthetic data distribution g easily, the only term in Eq. 2 need to learn is the importance weight $\mathrm{p}(X)/\mathrm{g}(X)$, defined by Eq. 5. Hence, an unbiased discriminator output $\Pr(X \sim \mathrm{p}|X)$ yields unbiased importance weights, and further, unbiased generalization loss.

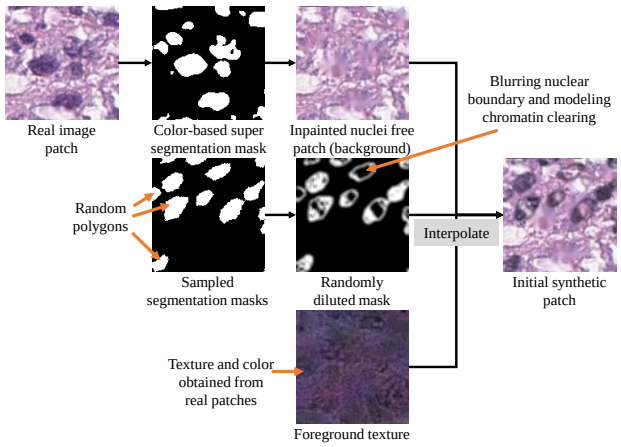


Figure 3. Inside our "GAN-free module": synthesizing a histopathology image patch utilizing textures in any given tissue type. This step generates an image patch matches the given mask.

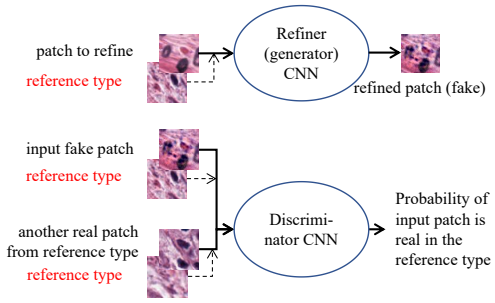


Figure 4. Inside our "GAN module": in addition to the input real/fake patches, we provide additional "reference type" patches extracted from nearby regions of the real patches. If the fake patch is realistic, but does not reflect the same tissue type as the reference type, the discriminator is still able to tell the difference. As a result, the refiner learns to generate patches in the reference style.

4. Heterogeneous Patch Synthesis

We now show how to synthesize (sample) training examples. Fig. 2 shows the overview of our method which learns from unlabeled real histopathology images of heterogeneous texture and cellular structure (*e.g.* nuclear mask).

4.1. Initial synthesis

This step generates synthetic patches that are not necessarily realistic for all given target tissue types. Thus, a significant part of this process is predefined regardless of

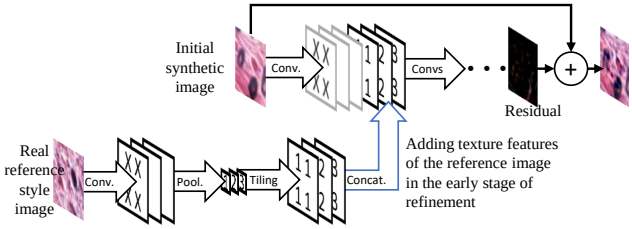


Figure 5. Our refiner (generator) CNN adds information of the reference type patch into the refinement stage, so that the initial synthetic patch will be refined according to the reference type.

the target tissue type. First, we randomly generate a set of polygons as nuclear masks. In particular, we perturb points on a circle closer/further away from the center according to a random irregularity value. These polygons are of variable sizes and irregularities and are allowed to randomly overlap with each other by a predefined number of pixels. To model the correlation between the shapes of nearby nuclei, all polygons are distorted by a random quadrilateral transform. The purpose of such a mask is to provide a generic representation of the basic structures in tissues and to induce greater variability in the synthetic images. We consider the generated masks as foreground/background masks (nuclei as the foreground and tissue as the background) and utilize textures from real histopathology image patches to generate initial synthetic image patches in a background/foreground manner. This is a fast process; synthesizing a 200×200 pixel patch at 40X magnification takes one second using a single CPU core.

Generating Background Patches: First, we remove the nuclei in a source image patch to create a background patch on which we add the synthetic nuclei. We apply a simple Ostu's threshold-based *super-segmentation* method [33] on the source image patch to determine the nuclear material. In super-segmentation, a segmented region always fully contains the foreground object (nucleus in this case). We replace the pixels corresponding to the segmented nuclear material with color and texture values similar to the background pixels via image inpainting [55]. Super-segmentation may not precisely delineate nucleus boundaries and may include non-nuclear material in segmented nuclei. This is acceptable, because the objective of this step is to guarantee that only background tissue texture and intensity properties are used to synthesize the background patch.

Simulating Foreground Nuclear Textures: We apply a *sub-segmentation* method to the source patch to gather nuclear textures from segmented regions. In sub-segmentation, a segmented region is fully contained in the foreground object. This ensures that pixels within real nuclei are used for generating realistic foreground (nuclei) in synthetic images. Since nuclei are generally small and make

up a small portion of tissue, sub-segmentation will yield a very limited amount of nuclear material which is not enough for existing reconstruction methods. Thus, our approach utilizes textures in the Eosin channel [19] of a randomly extracted real patch and combines them with nuclear color obtained via sub-segmentation of the source patch to generate nuclear textures.

Combining Foreground and Background: Let us define $I_{i,j}$, $A_{i,j}$, $B_{i,j}$, $M_{i,j}$ as pixel values at position i, j in the resulting synthetic patch, the nuclear texture patch, the nucleus free patch, and the nucleus mask patch, respectively. To combine nuclear and non-nuclear textures according to the nucleus mask patch, $I_{i,j}$ can be set to $A_{i,j}M_{i,j} + B_{i,j}(1 - M_{i,j})$. This may result in significant artifacts, such as obvious nuclear boundaries. Additionally, clear chromatin phenomena in certain types of nuclei are not modeled. Thus, our method randomly clears the interior of the polygons in the nucleus mask patch and blurs their boundaries before applying the above equation.

4.2. Refining the Initial Synthesis

These initial synthetic image patches are refined via adversarial training. We also use the discriminator's output to compute the importance sampling weight defined by Eq. 5. For this phase we have implemented a refiner (generator) CNN and a discriminator CNN.

Given an input image patch I and a reference type patch S, the refiner G with trainable parameters θ_G outputs a refined patch $X = \mathrm{G}(I, S; \theta_G)$. Ideally, an output patch is (1). *Regularized*: The pixel-wise difference between the initial synthetic patch and the refined patch is small enough so that the synthetic "ground truth" remains unchanged. (2). *Realistic for the given type*: It is a realistic representation of the type of the reference patch. (3). *Informative and hard*: It provides a challenging example for the task-specific CNN so that the trained task-specific CNN will be robust.

We construct three losses: L_G^{reg}, L_G^{real}, and L_G^{hard} for each of the properties above, respectively. The first two losses, L_G^{reg} and L_G^{real}, are based on the S+U method [51]. The weighted average of these losses is defined as the final loss L_G for training the refiner CNN:

$$L_G = \alpha L_G^{\text{reg}} + \beta L_G^{\text{real}} + \gamma L_G^{\text{hard}}. \tag{6}$$

We set hyperparameters $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.0000001$ in experiments.

The regularization loss L_G^{reg} is defined as an elastic net [70]: $L_G^{\text{reg}}(\theta_G) = \mathrm{E}\big[\lambda_1 ||I - X||_1 + \lambda_2 ||I - X||_2\big]$, where $\mathrm{E}[\cdot]$ is the expectation function applied on the training set, $||\cdot||_1$ and $||\cdot||_2$ are the L-1 and L-2 norms and λ_1 and λ_2 are predefined parameters. We use $\lambda_1 = 0.00001$ and $\lambda_2 = 0.0001$ in experiments.

The loss for achieving a realistic representation in the reference type, by training the refiner (generator) G, is $L_G^{\text{real}}(\theta_G) = \mathrm{E}\big[\log\big(1-\mathrm{D}(X, S; \theta_D)\big)\big]$, where $\mathrm{D}(X, S; \theta_D)$ is the output of the discriminator D with trainable parameters θ_D given the refined patch X and the same reference type patch S as input. It is the estimated probability by D that input X matches the tissue type of S. The discriminator D has two classes of input: pairs of real patches within the same type $\langle S', S\rangle$ and a pair with one synthetic patch $\langle X, S\rangle$. Its loss is the standard classification loss $L_D(\theta_D) = -\mathrm{E}\big[\log\big(\mathrm{D}(S', S; \theta_D)\big)\big] - \mathrm{E}\big[\log\big(1 - \mathrm{D}(X, S; \theta_D)\big)\big]$.

The generator and discriminator both take a reference patch and refine or classify the other input patch according to textures in the reference patch. This feature is implemented with an asymmetric siamese network [10, 28], as shown in Fig. 4 and Fig. 5.

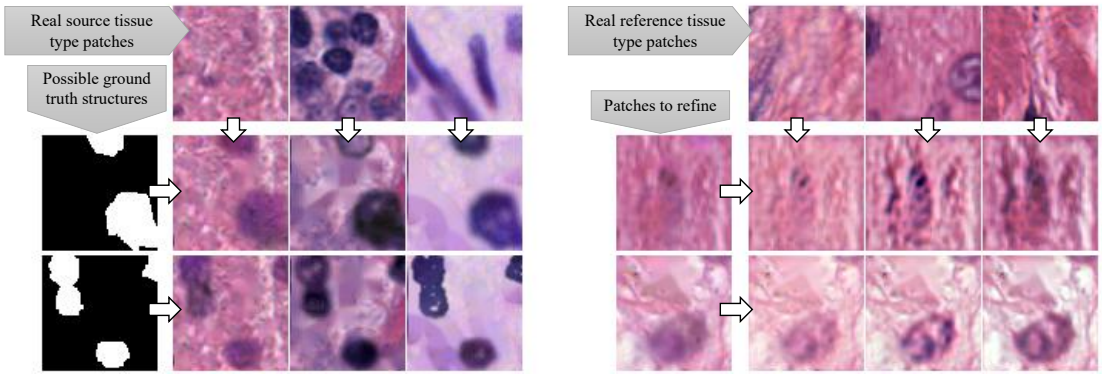
It has been shown that GANs are able to generate challenging training examples that yield robust classification/segmentation models [30, 50, 31, 21, 60]. Thus, the refiner is trained with loss L_G^{hard} to generate challenging training examples (with larger loss) for the task-specific CNN. We simply define L_G^{hard} as the negative of the task-specific loss: $L_G^{\text{hard}}(\theta_G) = -L_R(\theta_R)$, where $L_R(\theta_R)$ is the loss of a task-specific model R with trainable parameters θ_R. When training the refiner, we update θ_G to produce refined patches that maximize L_R. When training the task-specific CNN, we update θ_R to minimize L_R. The underlying segmentation ground truth of the refined patches would change significantly if $L_G^{\text{hard}}(\theta_G)$ overpowered $L_G^{\text{reg}}(\theta_G)$. We downweight L_G^{hard} by a factor of 0.0001 to minimize the likelihood of this unwanted outcome.

4.3. Visual Evaluation by a Human Expert

Fig. 6, 7, 8 show examples of our initial synthetic and refined patches. To verify that synthetic patches are realistic, we asked a pathologist to distinguish real versus synthetic patches. In particular, we showed the pathologist 100 randomly extracted real patches, 100 randomly selected initial synthetic patches, and 100 randomly selected refined patches. Out of this set, the pathologist selected the patches he thought were real. The pathologist classified almost half of the initial synthetic patches (**46%**) and most of the refined patches (**64%**) as real. The pathologist classified (**83%**) of the real patches as real. This is because many of those real patches are out-of-focus or contain no nuclei. Fig. 7 shows the distributions of weights of the realistic synthetic patches versus the unrealistic synthetic patches. This verifies that the realistic synthetic patches have higher importance sampling weights and vice versa.

5. Experiments

We conducted experiments with datasets from the MICCAI18 and MICCAI17 nucleus segmentation challenges [39, 58] and the generalized nucleus segmentation dataset



(a). Utilizing textures in various tissue types during the initial synthesis step.

(b). Refining the initial synthetic patches to be more like the reference type.

Figure 6. The effect of using different source tissue texture patches and reference type patches. The resulting synthetic patches have the same textures/types as the source/reference patches.

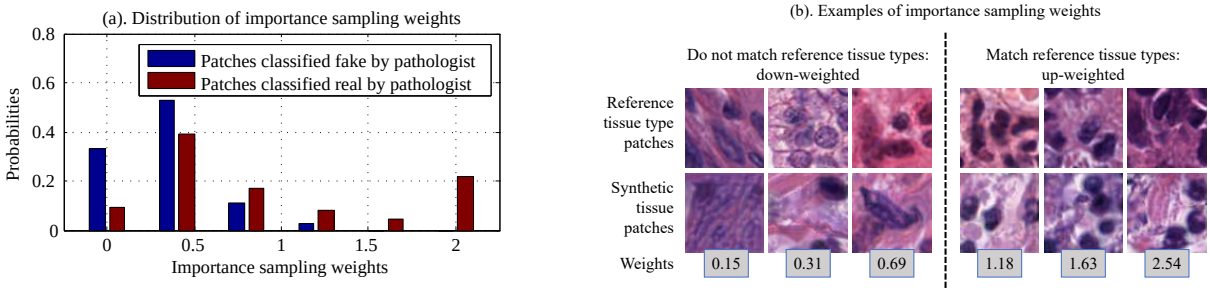


Figure 7. Evaluation and visualization of importance sampling weights. **(a)**. Synthetic patches classified as real by pathologists have higher importance weights than patches classified as fake. **(b)**. Visualization of importance sampling weights.

[29] containing seven cancer types. Additionally, we evaluated our method with a lymphocyte detection dataset [23].

We implemented the refiner, outlined in Fig. 5, with 21 convolutional layers and 2 pooling layers. The discriminator has 15 convolutional layers and 3 pooling layers. As the task-specific CNNs, we used U-net [47] and a network with 15 convolutional layers and 2 pooling layers for nucleus detection and segmentation, and a network with 11 convolutional layers for classification. For details, please refer to our source code. We used an open source implementation of GAN [27, 51] as part of our implementation. We initialize all networks randomly (no pretraining). During testing, we normalize the color of input H&E patches [45].

5.1. Nucleus Segmentation Experiments

Supervised methods heavily depend on representative datasets. However, currently only a few cancer types have supervised datasets due to the extensive amount of labor and expert domain knowledge required for histopathology image annotation. For cancer types without labeled data, supervised methods achieve worse performance than on cancer types with labeled data. We verified this argument using the MICCAI18 and MICCAI16/17 nucleus segmentation datasets [39, 58]. The MICCAI18 nucleus segmentation challenge dataset [39] contains 15 training and 18 testing tissue images extracted from whole slide images of two cancer types. The MICCAI17 dataset [58] contains 32 training and 32 testing images, extracted from whole slide images of four cancer types. A typical resolution is 600×600 pixels. In addition, we tested the across dataset generalization ability of our method using the test set of the generalized nucleus segmentation dataset [29]. The test set contains 14 1000×1000 pixel patches in seven cancer types.

Note that annotating one nucleus takes about 2 minutes. It would take about 225 man-hours to generate these training datasets. Unsupervised synthetic image generation and training can result in significant time savings in such cases, while enabling the generation of larger training datasets.

We evaluated several methods in the nucleus segmentation experiments; these methods are listed below. In the following, *Universal* denotes the proposed method trained with patches extracted from whole slide images for all cancer types in the TCGA repository. More specifically, we randomly extracted a 500×500-pixel tissue patch at 40X (for 20X images, we upsampled the patch to 40X) from each diagnostic whole slide image in the TCGA repository. This generated about 10k tissue patches.

Universal U-net. The proposed method with U-net [47]

as the task-specific CNN. Our U-net has two outputs: one for nucleus detection, and one for class-level nucleus segmentation. We then combined detection and class-level segmentation results to achieve instance-level segmentation using watershed [5, 2].

Universal CNN. The proposed method with a 15 layer segmentation/detection network.

Universal U-net + real data. Since U-net is computationally efficient, we train a U-net with both synthetic as well as real data from the MICCAI18 training dataset, as the model we deploy on over 5000 WSIs.

Type-specific U-net / CNN. We use the semi-supervised U-nets [47] and the 15/11 layer CNN as standalone **supervised** networks, trained with real, human annotated tissue image patches from up to four cancer types. We augment the real patches by rotation, mirroring, and scaling.

In order to obtain every tissue type for unsupervised learning of our method, we synthesized 75×75-pixel and 200×200-pixel patches according to patches sampled from every TCGA WSI. The "GAN-free module" generated 100k initial synthetic patches. Then we used GAN for image refinement and importance sampling based task-specific training on those initial synthetic patches.

We tested the supervised methods with the following two setups: **(1) Within cancer type.** We trained the type-specific, supervised CNNs with the training sets of all two MICCAI18 and four MICCAI17 cancer types. **(2) Across cancer types.** We excluded the training images of one cancer type, trained a type-specific, supervised CNN with the training images from all of the other cancer types, and evaluated the trained CNN on the images of the excluded type. We repeated this for all two/four cancer types and report performance as the average of all runs.

We used the average of two definitions of DICE coefficients as the performance metric. The first version is the standard DICE coefficient [18, 53]: denote the set of segmented pixels as S and the set of ground truth nuclear pixels as T, DICE$= 2 * |S \cap T|/(|S| + |T|)$. The second is a variant of the original to capture mismatch in the way the segmented objects are split, while the overall segmentation may be very similar. The evaluation results are shown in Tab. 1. Our approach outperforms the supervised methods significantly on testing cancer types without supervised data (across cancer types). Even when supervised data exists for every cancer type (within cancer type), our approach performs as well as the state-of-the-art approaches.

To further verify that our method outperforms baseline methods on tissue types without supervised data, we evaluated nucleus segmentation methods **across datasets**: we trained supervised method on the MICCAI17 training set and tested it on the test set of the generalized nucleus segmentation dataset [29]. As shown in Tab. 2, our method generalizes significantly better across datasets, than the supervised, type-specific method. Thus, we release segmentation results on 5000 WSIs in the TCGA repository [56]. Existing largest human annotated dataset [29] contains 100 patches of size 1000×1000 pixels. The scale of our segmentation results are larger than 10M such patches. We believe that this large-scale dataset, even though not as accurately annotated, is a useful feature for future pathology image analysis research.

Nucleus segmentation methods	MICCAI18 DICE Avg.	MICCAI17 DICE Avg.
Supervised methods tested *within cancer types*		
Type-specific CNN	0.8013	0.7713
Type-specific U-net	0.8391	0.7645
Contour-aware net [9]	0.812	-
CSP-CNN [23]	0.8362	0.7681
MICCAI18 winner	**0.870**	-
MICCAI17 winner [58]	-	0.783
Supervised methods tested *across cancer types*		
Type-specific CNN	0.7818	0.7314
Type-specific U-net	0.8010	0.7179
Proposed unsupervised method for *all cancer types*		
Universal CNN	0.8180	0.7708
Universal U-net	0.8401	0.7612
Universal U-net + real data	**0.8678**	**0.7863**

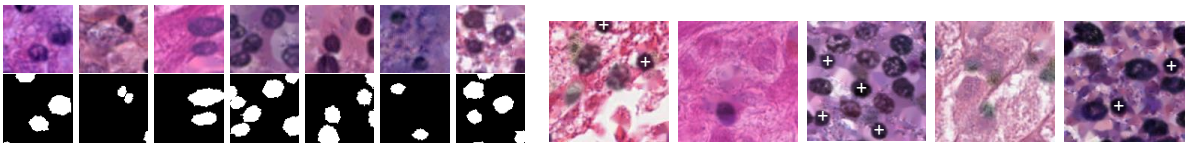
Table 1. Nucleus segmentation results on the MICCAI18 and MICCAI17 nucleus segmentation datasets. For each of the three network architecture, our approach outperforms the supervised methods significantly on cancer types without supervised data (across cancer). Even when supervised data exists for all cancer types (within cancer), our approach performs as well as state-of-the-art approaches without any supervision cost, due to the large scale of the synthetic dataset. The MICCAI18 winner's approach is unknown to us.

Nucleus segmentation methods	DICE Avg.
Type-specific U-net, across dataset	0.7328
Universal U-net + real data	**0.7713**

Table 2. Across dataset evaluation results. The type-specific CNN is trained on the MICCAI17 training set amd evaluated on the test set of the generalized nucleus segmentation dataset [29]. Our unsupervised method generalizes significantly better, than the supervised type-specific method.

5.2. Ablation Studies

We evaluated the importance of three components of our method: importance weights in the loss function, utilizing a real reference type patch for refinement, and generating hard examples for CNN training. We removed one feature at a time and measured performance for nucleus segmentation on the MICCAI17 dataset. The experimental results using U-net are shown in Tab. 3. The proposed methods reduce the segmentation error by 5.4%, 7.8%, and 3.2%.



(a). Synthetic examples for nucleus segmentation

(b). Synthetic examples for lymphocyte (indicated by +) detection

Figure 8. Examples of various kinds of synthetic patches we generated.

Nucleus segmentation methods	DICE Avg.
No hard examples	0.7476
No reference patch during refinement	0.7410
No importance weights	0.7533
Universal CNN (proposed)	**0.7612**

Table 3. Ablation study using the MICCAI17 nucleus segmentation challenge dataset. Proposed methods reduce the segmentation error (1−DICE average) by 5.4%, 7.8%, and 3.2%.

Lymphocyte detection methods	AUROC
Level Set features + supervised net [67]	0.7132
Fine-tuning VGG16 (supervised) [52]	0.6925
Universal CNN (proposed)	**0.7149**

Table 4. Lymphocyte detection on the lymphocyte dataset [23]. Without any supervision cost, our method outperforms all supervised models trained on patches of just one cancer type.

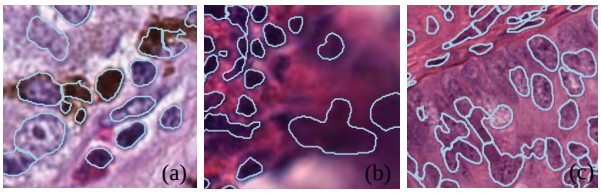


Figure 9. Three failure cases: Dark pigment in melanoma (a) and out-of-focus (b) scenarios are not modeled by our synthesis pipeline. Some light-colored nuclei with clear chromatin (c) are not detected when they are close to dark, easy-to-detect nuclei.

5.3. Human evaluation on 13 cancer types in TCGA

To evaluate nucleus segmentation methods in an uncontrolled environment, we randomly extracted 133 500×500 pixel patches from 13 major cancer types (that have more than 500 WSIs each) in TCGA [56], applied segmentation methods on those patches, and blindly compared the segmentation quality between our method and the baseline. For segmentation methods, we use the fully supervised U-net (type-specific U-net) trained on the MICCAI18 training set as the baseline, and the U-net trained on both synthetic and real MICCAI18 training data (Universal U-net + real data) as our method. For human evaluation, an expert pathologist blindly compared the segmentation results in terms of $TruePositives - FalsePositives - FalseNegatives$ in each patch. As a result, out of the 133 patches, in 83 patches our method is better than the baseline, in 46 patches our method is worse, in 4 patches they are similar. We show three failure cases in Fig. 9.

5.4. Lymphocyte Detection Experiments

The lymphocyte detection dataset [23] has 1367 labeled training patches and 418 testing patches cropped from 12 representative lung adenocarcinoma whole slide tissue images. Patches with lymphocytes in the center are labeled positive. Our method synthesized lymphocytes as round and dark objects with around 7 microns in diameter. Some synthetic image examples are shown in Fig. 8. Table 4 shows experimental evaluation of our method against a level set features based method [67] and supervised VGG16 method [52]. We used the Area Under the ROC curve (AUROC) measure as the evaluation metric.

6. Conclusions

Supervised methods rely on large volumes of labeled histopathology data which are expensive to generate. We introduced a method that learns from heterogeneous pathology patches in an unsupervised manner. Our method synthesizes training patches with importance weights, such that the task-specific (*e.g.* segmentation) CNN is trained to minimize the ideal (unbiased) generalization error over real data. When no supervised data exists for a cancer type, our result is significantly better than across-cancer generalization results by supervised methods. Even when supervised data exists, our approach performs as well as supervised methods, due to the much larger scale of synthetic data. We release segmentation results on over 5000 WSIs, which is orders of magnitude larger than currently available human annotated datasets. In future work we will demonstrate the generality of our importance sampling based loss minimization approach on other tasks such as mixed-quality image classification [63].

Acknowledgement This work was supported in part by 1U24CA180924-01A1, 3U24CA215109-02, and 1UG3CA225021-01 from the National Cancer Institute, R01LM011119-01 and R01LM009239 from the U.S. National Library of Medicine, and a gift from Adobe. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

References

[1] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 2014.

[2] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.

[3] N. Bayramoglu and J. Heikkilä. Transfer learning for cell nuclei classification in histopathology images. In *ECCV Workshops*, 2016.

[4] N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkila. Towards virtual h&e staining of hyperspectral lung histology images using conditional generative adversarial networks. In *CVPR*, 2017.

[5] S. Beucher. Watershed, hierarchical segmentation and waterfall algorithm. In *Mathematical morphology and its applications to image processing*. 1994.

[6] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Fulham. Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans). In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. 2017.

[7] C. M. Bishop. Pattern recognition and machine learning. 2006.

[8] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina. Biomedical data augmentation using generative adversarial neural networks. In *International Conference on Artificial Neural Networks*, 2017.

[9] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis*, 2017.

[10] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[11] R. Colen, I. Foster, R. Gatenby, M. E. Giger, R. Gillies, D. Gutman, M. Heller, R. Jain, A. Madabhushi, S. Madhavan, et al. Nci workshop report: clinical and computational requirements for correlating imaging phenotypes with genomics signatures. *Translational oncology*, 2014.

[12] F. S. Collins and H. Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 2015.

[13] L. A. Cooper, A. B. Carter, A. B. Farris, F. Wang, J. Kong, D. A. Gutman, P. Widener, T. C. Pan, S. R. Cholleti, A. Sharma, et al. Digital pathology: Data-intensive frontier in medical imaging. *Proceedings of the IEEE*, 2012.

[14] L. A. Cooper, J. Kong, D. A. Gutman, F. Wang, S. R. Cholleti, T. C. Pan, P. M. Widener, A. Sharma, T. Mikkelsen, A. E. Flanders, et al. An integrative approach for in silico glioma research. *IEEE Transactions on Biomedical Engineering*, 2010.

[15] L. A. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, S. Cholleti, T. Pan, A. Sharma, L. Scarpace, et al. Integrated morphologic analysis for the identification and characterization of disease subtypes. *Journal of the American Medical Informatics Association*, 2012.

[16] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho. Towards adversarial retinal image synthesis. *arXiv*, 2017.

[17] N. R. Council et al. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, 2011.

[18] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 1945.

[19] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 2008.

[20] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 2015.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[22] M. N. Gurcan and A. Madabhushi. Digital pathology. *SPIE*, 2013.

[23] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern Recognition*, 86:188–200, 2019.

[24] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, 2016.

[25] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. In *CVPR*, 2017.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[27] T. Kim. Simulated+unsupervised learning in tensorflow. https://github.com/carpedm20/simulated-unsupervised-tensorflow.

[28] G. Koch. Siamese neural networks for one-shot image recognition. In *ICML workshop*, 2015.

[29] 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 transactions on medical imaging*, 2017.

[30] H. Le, T. F. Y. Vicente, V. Nguyen, M. Hoai, and D. Samaras. A+ D net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, 2018.

[31] J. Lemley, S. Bazrafkan, and P. Corcoran. Smart augmentation-learning an optimal data augmentation strategy. *IEEE Access*, 2017.

[32] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.

[33] P.-S. Liao, T.-S. Chen, P.-C. Chung, et al. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, 2001.

[34] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[35] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[36] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[37] F. Mahmood, D. Borders, R. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *arXiv*, 2018.

[38] A. Mauricio, J. López, R. Huauya, and J. Diaz. High-resolution generative adversarial neural networks applied to histological images generation. In *International Conference on Artificial Neural Networks*, 2018.

[39] MICCAI 2018 Challenge. Segmentation of Nuclei in Pathology Images. http://miccai.cloudapp.net/competitions/83, 2018.

[40] V. Murthy, L. Hou, D. Samaras, T. M. Kurc, and J. H. Saltz. Center-focusing multi-task CNN with injected features for classification of glioma nuclear images. In *Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[41] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *CVPR*, 2015.

[42] A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi. Gans for biological image synthesis. In *ICCV*, 2017.

[43] C. Parmar, R. T. Leijenaar, P. Grossmann, E. R. Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin, and H. J. Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports*, 2015.

[44] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[45] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[46] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[47] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[48] J. Saltz, J. Almeida, Y. Gao, A. Sharma, E. Bremer, T. DiPrima, M. Saltz, J. Kalpathy-Cramer, and T. Kurc. Towards generation, management, and exploration of combined radiomics and pathomics datasets for cancer research. *AMIA Summits on Translational Science Proceedings*, 2017.

[49] S. Sankaran, L. R. Khot, C. Z. Espinoza, S. Jarolmasjed, V. R. Sathuvalli, G. J. Vandemark, P. N. Miklas, A. H. Carter, M. O. Pumphrey, N. R. Knowles, et al. Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review. *European Journal of Agronomy*, 2015.

[50] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[51] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014.

[53] T. Sørensen. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. Skr.*, 1948.

[54] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *ICLR*, 2017.

[55] A. Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 2004.

[56] The TCGA team. The Cancer Genome Atlas. https://cancergenome.nih.gov/.

[57] T. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, 2016.

[58] Q. D. Vu, S. Graham, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, T. Kurc, K. Farahani, T. Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *arXiv*, 2018.

[59] S. Wang, J. Yao, Z. Xu, and J. Huang. Subtype cell detection with an accelerated deep convolution neural network. In *MICCAI*, 2016.

[60] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017.

[61] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *MICCAI*, 2015.

[62] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *Medical Imaging*, 2016.

[63] F. Yang, Q. Zhang, M. Wang, and G. Qiu. Quality classified image analysis with application to face detection and recognition. *arXiv*, 2018.

[64] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *MICCAI*, 2017.

[65] C. Yuan, Y. Zhang, and Z. Liu. A survey on technologies for automatic forest fire monitoring, detection, and fighting using unmanned aerial vehicles and remote sensing techniques. *Canadian journal of forest research*, 2015.

[66] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 2017.

[67] J. Zhao, L. Xiong, K. Jayashree, J. Li, F. Zhao, Z. Wang, S. Pranata, S. Shen, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, 2017.

[68] N. Zhou, X. Yu, T. Zhao, S. Wen, F. Wang, W. Zhu, T. Kurc, A. Tannenbaum, J. Saltz, and Y. Gao. Evaluation of nucleus segmentation in digital pathology images through large scale image synthesis. *SPIE Medical Imaging. International Society for Optics and Photonics*, 2017.

[69] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.

[70] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.