

# Dataset of Segmented Nuclei in Hematoxylin and Eosin Stained Histopathology Images (Pan-Cancer-Nuclei-Seg)

## Summary

Detection, segmentation and classification of nuclei are fundamental analysis operations in digital pathology. Existing state-of-the-art approaches demand extensive amounts 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-weights 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.

In this dataset we release code and nucleus segmentations in whole slide tissue images with quality control results for over 5000 Whole Slide Images (WSI) in The Cancer Genome Atlas (TCGA) repository. There are two subsets of data: (1) automatic nucleus segmentation data of 5,060 whole slide tissue images of 10 cancer types, with quality control results, and (2) manual nucleus segmentation data of 1,356 image patches from the same 10 cancer types plus additional 4 cancer types.

**These 5,060 Whole Slide Images (WSIs) are from the following 10 cancer types:**

**BLCA** Bladder urothelial carcinoma

**BRCA** Breast invasive carcinoma

**CESC** Cervical squamous cell carcinoma and endocervical adenocarcinoma

**GBM** Glioblastoma Multiforme

**LUAD** Lung adenocarcinoma

**LUSC** Lung squamous cell carcinoma

**PAAD** Pancreatic adenocarcinoma

**PRAD** Prostate adenocarcinoma

**SKCM** Skin Cutaneous Melanoma

**UCEC** Uterine Corpus Endometrial Carcinoma

Note that you can also download segmentation data of following 4 cancer types, although they are not officially verified or released.

**COAD** Colon adenocarcinoma

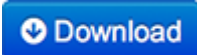
**READ** Rectal adenocarcinoma

**STAD** Stomach adenocarcinoma

**UVM** Uveal Melanoma

### Data Access

### Data Access

Data Type	Download all or Query/Filter
Tissue Slide Segmentation Results (1,200 GB)	

WSI groups	Percentage of patches	
	with bad segmentations	#. slides
Best	0%	2,346
Good	0.01 - 6.67%	1,246
Adequate	6.68 - 13.3%	593
Problematic	13.4 - 20.0%	302
Unacceptable	> 20.0% or failed WSI QC	573

Table 2: We categorize WSIs into groups with different segmentation quality levels. Slides identified as having unacceptable segmentation results are excluded from analysis in the rest of this work.

List of histopathology slides (TXT, 348.5 KB)	<a href="#">Download</a>
Whole slide image-level quality control results (TXT, 151.4 KB)	<a href="#">Download</a>
Segmentation region checking results (TXT, 169.4 KB)	<a href="#">Download</a>

Click the Versions tab for more info about data releases.

**Note:** Please contact [help@cancerimagingarchive.net](mailto:help@cancerimagingarchive.net) with any questions regarding usage.

#### Collections Used in this Third Party Analysis

Below is a list of the Collections used in these analyses:

- [TCGA-BLCA](#),
- [TCGA-BRCA](#),
- [TCGA-CESC](#),
- [TCGA-COAD](#),
- [TCGA-GBM](#),
- [TCGA-LUAD](#),
- [TCGA-LUSC](#),
- [TCGA-PAAD](#),
- [TCGA-PRAD](#),
- [TCGA-READ](#),
- [TCGA-SKCM](#),
- [TCGA-STAD](#),
- [TCGA-UCEC](#),
- [TCGA-UVM](#)

#### Detailed Description

### Detailed Description

Additional visual segmentation data can be found on [PathDB](#)

#### Manual nucleus segmentation data of 1,356 patches

These 1,356 patches are randomly extracted from all 14 cancer types mentioned above. This data contains original H&E stained histopathology image patches, and instance-level segmentation masks. Additional information is in the [readme.docx](#) file.

#### Citations & Data Usage Policy

### Citations & Data Usage Policy

Users of this data must abide by the [TCIA Data Usage Policy](#) and the [Creative Commons Attribution 3.0 Unported License](#) under which it has been published. Attribution should include references to the following citations:

#### Dataset Citation

Hou, L., Gupta, R., Van Arnam, J. S., Zhang, Y., Sivalenka, K., Samaras, D., Kurc, T., & Saltz, J. H. (2019). **Dataset of Segmented Nuclei in Hematoxylin and Eosin Stained Histopathology Images of 10 Cancer Types** [Data set]. The Cancer Imaging Archive. <https://doi.org/10.7937/TCIA.2019.4A4DKP9U>

#### Publication Citation

Hou, L., Agarwal, A., Samaras, D., Kurc, T. M., Gupta, R. R., & Saltz, J. H. (2019, June). **Robust Histopathology Image Analysis: To Label or to Synthesize?** 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8533-8542. <https://doi.org/10.1109/cvpr.2019.00873> [Open Access Here](#)

#### TCIA Citation




Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**. *Journal of Digital Imaging*, 26(6), 1045–1057. <https://doi.org/10.1007/s10278-013-9622-7>

### Other Publications Using This Data

TCIA maintains [a list of publications](#) that leverage TCIA data. If you have a manuscript you'd like to add please [contact the TCIA Helpdesk](#).

#### Versions

**Version 1 (Current): 2020/02/08**

Data Type	Download all or Query/Filter
Tissue Slide Images (SVS, 1,200 GB)	
List of histopathology slides (TXT, 348.5 KB )	
WSI quality control results (TXT, 151.4 KB)	
Segmentation region checking results (TXT, 169.4 KB)	