Read Me

This repository contains segmentation results computed by the method described in the paper below:

L. Hou, A. Agarwal, D. Samaras, T.M. Kurc, R.R. Gupta, J.H. Saltz. Robust Histopathology Image Analysis: to label or to Synthesize? Oral presentation in CVPR. 2019.

<https://www3.cs.stonybrook.edu/~cvl/content/papers/2019/Hou_CVPR19.pdf>

**List of cancer types**

We have generated nucleus segmentation results for the following cancer types:

|  |  |
| --- | --- |
| **Abbreviation** | **Cancer type** |
| BLCA | Urothelial carcinoma of the bladder |
| BRCA | Invasive carcinoma of the breast |
| CESC | Cervical squamous cell carcinoma and endocervical adenocarcinoma |
| COAD | Colon adenocarcinoma |
| GBM | Glioblastoma Multiforme |
| LUAD | Lung adenocarcinoma |
| LUSC | Lung squamous cell carcinoma |
| PAAD | Pancreatic adenocarcinoma |
| PRAD | Prostate adenocarcinoma |
| READ | Rectal adenocarcinoma |
| SKCM | Skin Cutaneous Melanoma |
| STAD | Stomach adenocarcinoma |
| UCEC | Endometrial Carcinoma of the Uterine Corpua |
| UVM | Uveal Melanoma |

All Whole Slide Images (WSI) are from The Cancer Genome Atlas (TCGA) repository (<https://cancergenome.nih.gov/>).

Based on our evaluation, segmentation results in READ, COAD, STAD, and UVM are suboptimal.

**Format of segmentation results**

For each cancer type, you can find three folders. For example, for BLCA:

|  |  |
| --- | --- |
| BLCA\_polygon | Contains polygon coordinates for each segmented nucleus, for all WSIs of BLCA.  These results are obtained by:   1. Thresholding grayscale results in the BLCA\_prob folder. 2. Nucleus declumping using the watershed algorithm. This declumping step combines the nucleus detection and segmentation results, to separate touching or overlapping nuclei. |
| BLCA\_meta | Contains meta data, for all WSIs of BLCA.  This is useless unless you use caMicroscope (<https://github.com/camicroscope/caMicroscope>) for visualizing WSIs and segmentation results. |

Inside each folder, you can find a tar.gz file for each WSI. The name of the tar.gz file is the same as the name of the WSI in TCGA. After uncompressing a tar.gz file, you can find many files. **Each file contains results in a large tile (4000 by 4000 pixels).** The filename encodes the position of the tile.

Take the following file for example:

**blca**\_**polygon**/

* **TCGA-4Z-AA7Q-01Z-00-DX1.9C30EAED-8DE3-437C-8852-0C64B415AFA8.svs**/
  + **16001**\_**48001**\_4000\_4000\_0.2525\_1-features.csv

**blca**: this slide is Urothelial carcinoma of the bladder (BLCA).

**polygon**: this folder contains the polygon coordinates for each segmented nucleus.

**TCGA-4Z-AA7Q-01Z-00-DX1.9C30EAED-8DE3-437C-8852-0C64B415AFA8.svs**: name of the BLCA slide.

**16001**: the x position (horizontal axis) of the tile's top left corner.

**48001**: the y position (vertical axis) of the tile's top left corner.

Also take the example above (16001\_48001\_4000\_4000\_0.2525\_1-features.csv), the content of the file is below:

AreaInPixels,PhysicalSize,Polygon

31,31,[17737.0:48001.0:17737.0:48006.0:17738.0:48006.0:17739.0:48007.0:17740.0:48007.0:...]

167,167,[19473.0:48001.0:19472.0:48002.0:19467.0:48002.0:19467.0:48003.0:19463.0:48007.0:...]

...

Each line in a csv file under XXX\_polygon/ contains information of one nucleus. There are three columns in a csv file:

|  |  |
| --- | --- |
| AreaInPixels | Size of the nucleus, in terms of the number of pixels, regardless of the magnification level of the Whole Slide Image (WSI). |
| PhysicalSize | Size of the nucleus, in terms of the projected number of pixels in 40X magnification level. |
| Polygon | We represent the contour of the nucleus by a polygon. Vertices of the polygon are saved in format [x0:y0:x1:y1:x2:y2.......]. |

Additional Data

|  |  |  |
| --- | --- | --- |
| **Data** | **Description** | **Meta-info** |
| manual\_segmentation\_data.tar.gz | Manual segmentation data on 1,356 patches of 256 × 256 pixels in 40X, uniformly sampled from 14 cancer types. | Indexed by cancer type, slide id, patch coordinate, ID of the human annotator. |
| random\_segmentation\_region\_checking\_result.txt | Random segmentation region checking result: records that shows for each WSI, how many patches are randomly sampled and how many patches have bad segmentation results. | Indexed by cancer type and slide id. |
| wsi\_quality\_control\_result.txt | The list of slides selected for WSI-level quality control by visual assessment. Slides with unacceptable results are marked. | Indexed by cancer type and slide id. |
| list\_of\_verified\_wsis.txt | The list of WSIs that we run a quality control process. These WSIs are from 10 cancer types (without STAD, COAD, READ, UVM). | Indexed by cancer type, slide id, and whether the slide passes quality control |

Other useful information

Please contact Le Hou (le.hou@stonybrook.edu) if you have questions.

Segmentation results for some slides might be missing due to technical errors, such as being unable to read whole slide images.

You can also find the source code and a trained model of our segmentation method online:

<https://github.com/SBU-BMI/quip_cnn_segmentation>

If you want to convert polygons to segmentation masks, check out this readme:

<https://github.com/SBU-BMI/quip_cnn_segmentation/blob/master/segmentation-of-nuclei/READMD.md#extracting-segmentation-mask-from-output-folder>