A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi

Supplementary Material

1. Information about the TCGA images used in this work

Table S1: *Information about the images used in this paper including their organ, disease type, tissue source site codes and hospital names where the tissue sample was prepared is given below. Tissue sample of patient no. 21 (highlighted in red) was used as a target image for color normalization. To know more about TCGA patient codes and hospital details, the readers can consult TCGA website*[1].

#	Patient ID	Organ	Disease type	Tissue Source Site Code	Hospital/Clinic
1	TCGA-A7-A13E-01Z-00-DX1	Breast	Breast invasive carcinoma	A7	Christiana Healthcare
2	TCGA-A7-A13F-01Z-00-DX1	Breast	Breast invasive carcinoma	A7	Christiana Healthcare
3	TCGA-AR-A1AK-01Z-00-DX1	Breast	Breast invasive carcinoma	AR	Mayo Clinic
4	TCGA-AR-A1AS-01Z-00-DX1	Breast	Breast invasive carcinoma	E2	Roswell Park
5	TCGA-E2-A1B5-01Z-00-DX1	Breast	Breast invasive carcinoma	E2	Roswell Park
6	TCGA-E2-A14V-01Z-00-DX1	Breast	Breast invasive carcinoma	E2	Roswell Park
7	TCGA-B0-5711-01Z-00-DX1	Kidney	Kidney renal clear cell carcinoma	B0	University of Pittsburgh
8	TCGA-HE-7128-01Z-00-DX1	Kidney	Kidney renal papillary cell carcinoma	HE	Ontario Institute for Cancer Research (OICR)
9	TCGA-HE-7129-01Z-00-DX1	Kidney	Kidney renal papillary cell carcinoma	HE	Ontario Institute for Cancer Research (OICR)
10	TCGA-HE-7130-01Z-00-DX1	Kidney	Kidney renal papillary cell carcinoma	HE	Ontario Institute for Cancer Research (OICR)
11	TCGA-B0-5710-01Z-00-DX1	Kidney	Kidney renal clear cell carcinoma	B0	University of Pittsburgh
12	TCGA-B0-5698-01Z-00-DX1	Kidney	Kidney renal clear cell carcinoma	B0	University of Pittsburgh
13	TCGA-18-5592-01Z-00-DX1	Liver	Lung squamous cell carcinoma	18	Princess Margaret Hospital (Canada)
14	TCGA-38-6178-01Z-00-DX1	Liver	Lung adenocarcinoma	38	University of North Carolina
15	TCGA-49-4488-01Z-00-DX1	Liver	Lung adenocarcinoma	49	Johns Hopkins
16	TCGA-50-5931-	Liver	Lung	50	University of

[1] https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode

	01Z-00-DX1		adenocarcinoma		Pittsburgh
17	TCGA-21-5784-01Z-00-DX1	Liver	Lung squamous cell carcinoma	21	Fox Chase Cancer Center
18	TCGA-21-5786-01Z-00-DX1	Liver	Lung squamous cell carcinoma	21	Fox Chase Cancer Center
19	TCGA-G9-6336-01Z-00-DX1	Prostate	Prostate adenocarcinoma	G9	Roswell Park
20	TCGA-G9-6348-01Z-00-DX1	Prostate	Prostate adenocarcinoma	G9	Roswell Park
21	**TCGA-G9-6356-01Z-00-DX1**	**Prostate**	**Prostate adenocarcinoma**	**G9**	**Roswell Park**
22	TCGA-G9-6363-01Z-00-DX1	Prostate	Prostate adenocarcinoma	G9	Roswell Park
23	TCGA-CH-5767-01Z-00-DX1	Prostate	Prostate adenocarcinoma	CH	Indivumed
24	TCGA-G9-6362-01Z-00-DX1	Prostate	Prostate adenocarcinoma	G9	Roswell Park
25	TCGA-DK-A2I6-01A-01-TS1	Bladder	Bladder Urothelial Carcinoma	DK	Memorial Sloan Kettering
26	TCGA-G2-A2EK-01A-02-TSB	Bladder	Bladder Urothelial Carcinoma	G2	MD Anderson
27	TCGA-AY-A8YK-01A-01-TS1	Colon	Colon adenocarcinoma	AY	University of North Carolina
28	TCGA-NH-A8F7-01A-01-TS1	Colon	Colon adenocarcinoma	NH	Candler
29	TCGA-KB-A93J-01A-01-TS1	Stomach	Stomach adenocarcinoma	KB	University Health Network, Toronto
30	TCGA-RD-A8N9-01A-01-TS1	Stomach	Stomach adenocarcinoma	RD	Peter MacCallum Cancer Center

2. Close-up examples of under-segmentation and over-segmentation

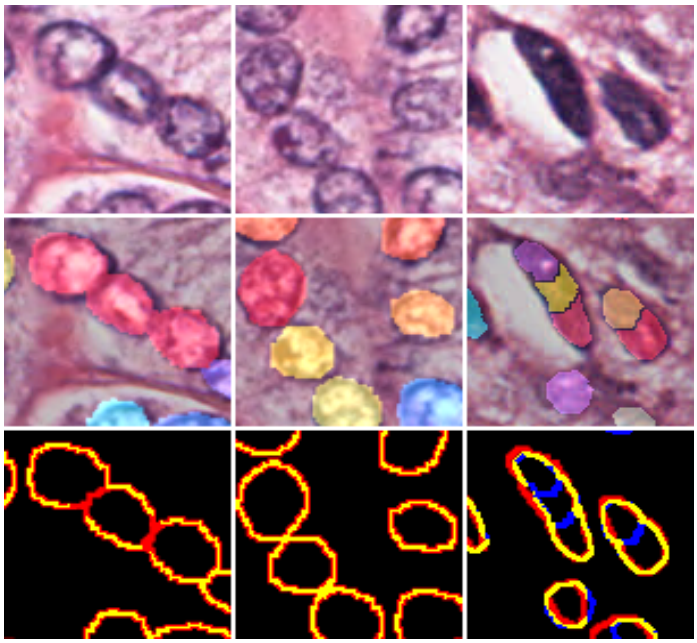

Figure S1: Columns illustrate **(a)** under-segmentation, **(b)** correct segmentation, and **(c)** over-segmentation. The top row shows the raw image and the middle row shows (predicted) overlaid nuclear segments (each with a separate color). In the bottom row, red curves represent ground truth boundary, blue represent predicted boundary and yellow represents pixels

where the ground truth and detected boundaries overlap. Notice the predominance of yellow curves (correct segmentation) in **(b)**, red (undetected true boundaries) in **(a)**, and blue (falsely detected boundaries) in **(c)**.

Throughout the paper, we use the term under-segmentation for a segment that is unduly large. That is, it almost completely covers a ground truth nucleus and additionally covers significant area outside that nucleus, which may include neighboring nuclei as illustrated in Figure S1 (a). We consider correct segmentation as the one where the segmented nuclear boundary (and thus shape) exactly overlaps with the ground truth nuclear boundary as shown in Figure S1 (b). Similarly, when most of the pixels of an unduly small segment belong mostly to only one nucleus but it leaves out a large proportion of other pixels from the same nucleus uncovered, we call it over-segmentation. This also includes cases where a nucleus is split into multiple detected objects as illustrated in Figure S1 (c).

3. Protocol used for the evaluation of the annotation quality by an expert pathologist

We sent annotated images to an expert pathologist for examination of annotation quality. In a PowerPoint® deck, we used one image per slide. On a slide, we put the unannotated and annotated images side by side to cover a large portion of the slide. The pathologist viewed the slide on a 25" monitor, and was instructed to place an arrow shape on every problematic annotation, whether it was a false positive, a false negative, an over-segmented, or an under-segmented nucleus. Examples of the pathologist's assessment are shown in Figure S2. Each of the sub-figures covered almost a complete PowerPoint® slide.

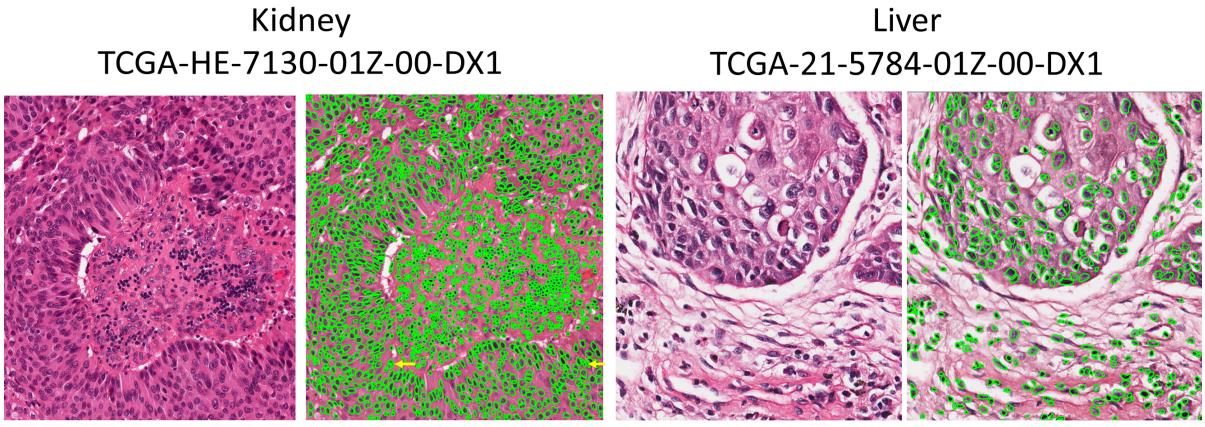

Figure S2 (a) Pathologist's Comment: "Tough due to extreme variation in nuclear size and shape – but good overall. Note over-segmentation (yellow arrows)." **(b) Pathologist's Comment**: "Very good recognition of overlapping nuclei. This is an unusual and difficult image."

We counted all the arrows and divided the count by the number of annotated nuclei in those images to estimate that our annotators made less than 1% errors on any given image. The total number of annotated nuclei and the annotation errors pointed by the expert pathologist for each tissue type are given in Table S2. We left these errors uncorrected due to their low count as evident from Table S2.

Table S2: *Variation of the annotation errors with the tissue (or organ) type*

#	Organ	# of nuclei annotated	# of erroneous annotations
1	Breast	3,712	21
2	Liver	4,593	18
3	Kidney	7,228	47
4	Prostate	2,485	23
5	Bladder	768	4
6	Colon	742	6
7	Stomach	2,617	19
Total		**22,145**	**138**