DeepLIIF: Deep Learning-Inferred Multiplex ImmunoFluorescence for IHC Quantification

Parmida Ghahremani[1], Yanyun Li[3], Arie Kaufman[1], Rami Vanguri[3], Noah Greenwald[2], Michael Angelo[2], Travis J. Hollmann[3,*], and Saad Nadeem[3,*,✉]

[1]Stony Brook University, Stony Brook, NY, USA
[2]Stanford University, Stanford, CA, USA
[3]Memorial Sloan Kettering Cancer Center, New York, NY, USA

Reporting biomarkers assessed by routine immunohistochemical (IHC) staining of tissue is broadly used in diagnostic pathology laboratories for patient care. To date, clinical reporting is predominantly qualitative or semi-quantitative. By creating a multitask deep learning framework referred to as DeepLIIF, we are presenting a single step solution to nuclear segmentation and quantitative single-cell IHC scoring. Leveraging a unique *de novo* dataset of co-registered IHC and multiplex immunoflourescence (mpIF) data generated from the same tissue section, we simultaneously segment and translate low-cost and prevalent IHC slides to more expensive-yet-informative mpIF images. Moreover, a nuclear-pore marker, LAP2beta, is co-registered to improve cell segmentation and protein expression quantification on IHC slides. By formulating the IHC quantification as cell instance segmentation/classification rather than cell detection problem, we show that our model trained on clean IHC Ki67 data can generalize to more noisy and artifact-ridden images as well as other nuclear and non-nuclear markers such as CD3, CD8, BCL2, BCL6, MYC, MUM1, CD10 and TP53. We thoroughly evaluate our method on publicly available benchmark datasets as well as against pathologists' semi-quantitative scoring. The code, trained models, and the resultant embeddings for all the datasets used in this paper will be released via GitHub.

Mulitask Learning | Multiplex ImmunoFluoresence | Immunohistochemisty

*** Co-senior authors**
(✉) Correspondence: *nadeems@mskcc.org*

Introduction

The assessment of protein expression using immunohistochemical staining of tissue sections on glass slides is critical for guiding clinical decision making in several diagnostic clinical scenarios including cancer classification, residual disease detection and even mutation detection (BRAFV600E and NRASQ61R). The conventional method of assessment is manual semi-quantitative ('positive', 'negative', 'low', 'medium', 'high' or approximate percentage staining within a population) scoring of different proteins by an anatomic pathologist after tissue staining by immunohistochemistry (IHC). Standard chromogenic IHC staining, while high throughput, has a narrow dynamic range and a relatively limited number of markers are detectable on the same slide. The restricted marker depth or "plexing" of standard IHC limits further delineation of which cells are expressing the protein-of-interest (Ki67, PDL1, Bcl6, etc). Furthermore, the limited marker depth of IHC prevents the inclusion of markers of cell boundaries and therefore manual cell segmentation is also highly error prone with high inter-observer variability.

As opposed to conventional immunohistochemistry (IHC) staining, multiplex immunofluorescence (mpIF) staining provides the opportunity to examine panels of several markers individually or simultaneously as a composite permitting accurate co-localization, stain standardization, more objective scoring, and cut-offs for all the markers values (especially in low-expression regions, which are difficult to assess on IHC stained slides and can be misconstrued as negative due to weak staining that can be masked by the hematoxylin counterstain) (1, 2). Moreover, in a recent meta-analysis (3), mpIF was shown to have a higher diagnostic prediction accuracy (at par with multimodal cross-platform composite approaches) than IHC scoring, tumor mutational burden, or gene expression profiling. However, mpIF assays are expensive and not widely available. This can lead to a unique opportunity to leverage the advantages of mpIF to improve the explainability and interpretability of the conventional IHCs using recent deep learning breakthroughs.

Current deep learning methods for scoring IHCs rely solely on the error-prone manual annotations (unclear cell boundaries, overlapping cells, and difficult assessment of low-expression regions) rather than on co-registered high-dimensional imaging of the same tissue samples. Therefore, we present a new multitask deep learning technique, leveraging co-registered IHC and mpIF data for different tissue and cancer types, to simultaneously translate low-cost/prevalent IHC images to high-cost and more informative mpIF representations (creating a Deep-Learning-Inferred IF image), accurately auto-segment relevant cells, and quantify protein expression for more accurate and reproducible IHC quantification; using multitask learning (4) to train models to perform a variety of tasks rather than one narrowly defined task makes them more generally useful and robust. *In essence, this creates registered orthogonal datasets to confirm and further specify the target staining characteristics. The benefit of our model is that we establish an absolute and quantitative single-cell IHC scoring system rather than the semi-quantitative/binned criteria often used clinically.*

Several approaches have been proposed for deep learning-based stain-to-stain translation of unstained (label-free), H&E, IHC and multiplex slides but *relatively few attempts have been made (in limited contexts) at leveraging the translated enriched feature set for cellular-level segmentation,*

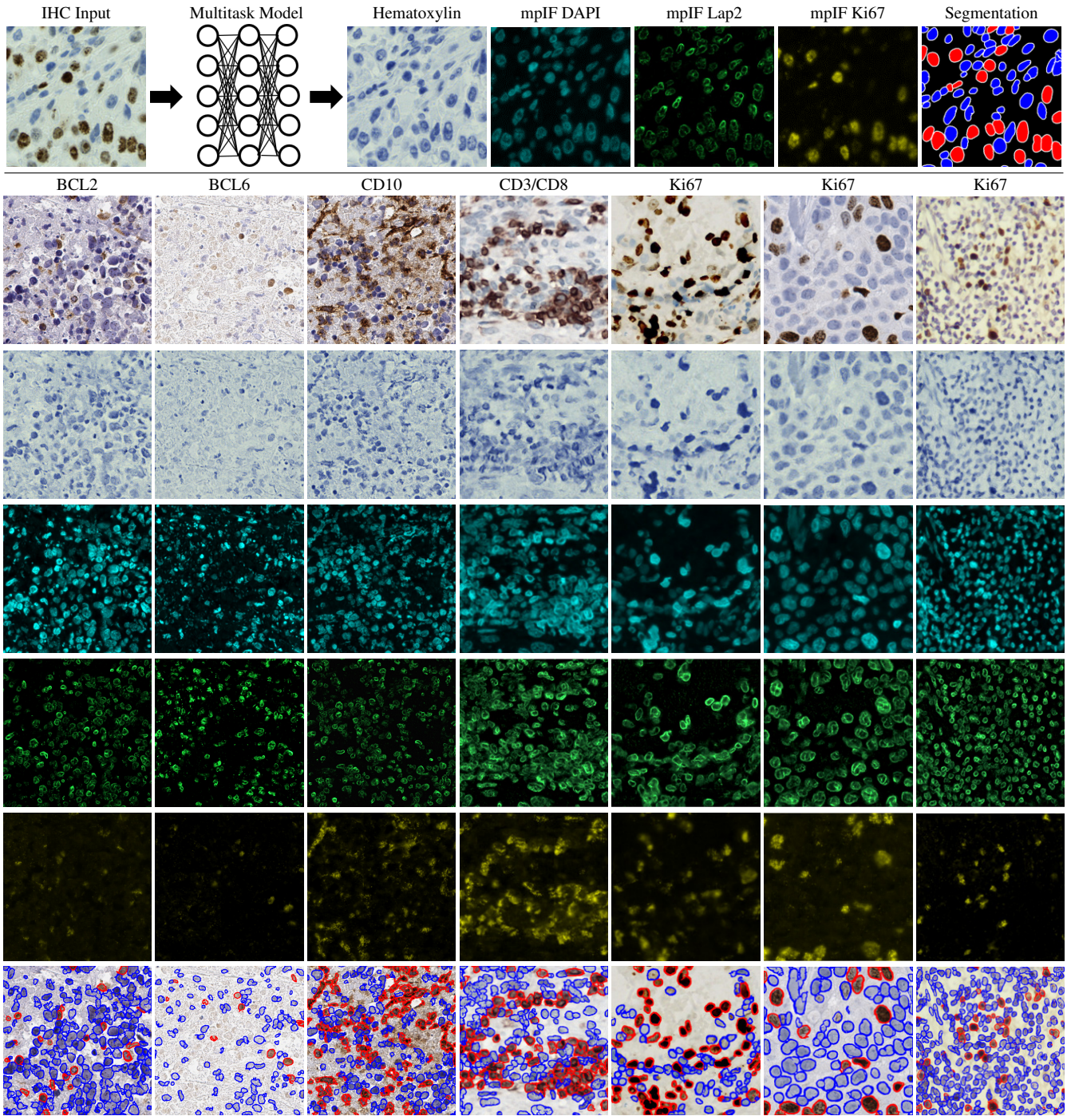


Fig. 1. DeepLIIF multitask deep learning results for IHC Ki67. Given IHC input, our multitask deep learning framework simultaneously generates the corresponding IHC Hematoxylin channel, mpIF DAPI, mpIF protein expression (Ki67, CD3, CD8, etc.) and the nuclei segmentation, baking explainability and interpretability into the model itself rather than relying on coarse activation/attention maps. In the segmentation mask, the red color shows positive cells while the blue color shows the negative cells.

classification or scoring (5, 6). Recently, Liu et al. (7) used publicly available fluorescence microscopy and histopathology H&E datasets for unsupervised nuclei segmentation in histopathology images, by learning from fluorescence microscopy DAPI images. However, their pipeline incorporated CycleGAN which hallucinated nuclei in the target histopathology domain and hence, required segmentation masks in the source domain to remove any redundant or unnecessary nuclei in the target domain. The model was also not generalizable across the two target histopathology datasets due to the stain variations, making this unsupervised solution less suitable for inferring different cell types from given H&E or IHC images. Burlingame et al. (8) on the other hand used supervised learning trained on H&E and co-registered single-channel pancytokeratin IF for 4 pancreatic ductal adenocarcinoma (PDAC) patients to infer pancytokeratin stain for given PDAC H&E image. Another work (9) used a supervised deep learning method trained on H&E and

co-registered IHC PHH3 DAB slides for mitosis detection in H&E breast cancer WSIs. Moreover, for stain-to-stain translation, there are methods to translate between H&E and IHC but none for translating between IHC and mpIF modalities. *To focus on immediate clinical application, we want to accentuate/disambiguate the cellular information in low-cost IHCs (using a higher-cost and more informative mpIF representation) to improve the interpretability for pathologists as well as for the downstream analysis/algorithms.* Traditional IHC deconvolution or stain separation algorithms do not work well in our context and are difficult to generalize even across the same patient cohort.

In recent years, deep convolutional neural networks have achieved great success in the automatic analysis of medical images, including nuclei detection, segmentation, and classification in digital pathology images. Long et al. (10) designed the fully convolutional neural network (FCN) for semantic segmentation. Several other popular FCN-based architectures such as SegNet (11), DeepLab (12), RefineNet (13, 14) achieved state-of-the-art performance. Ronneberger et al. (15) proposed U-Net, an FCN-based network architecture to detect nuclei from the background by utilizing features from different scales. However, this network usually fails in separating touching and overlapping nuclei. Zhou et al. (16) presented UNet++ for reducing the semantic gap between the feature maps of the encoder and decoder of the UNet by adding a series of nested, dense skip pathways, resulting in higher accuracy in image segmentation tasks in comparison with UNet. He et al. (17) achieved higher accuracy on various semantic segmentation tasks by designing Mask_RCNN, a Region-based CNN (RCNN) approach. This model generates three outputs for each candidate object, a class label, a bounding-box offset, and an object mask, and performs pixel-to-pixel alignment, which makes it a powerful segmentation model. Several cell counting approaches are designed specifically for detecting the centroids of the cells, using a cell spatial density mask (18–20). These approaches, however, assume that the cells have circular morphology, resulting in their failure to detect cells with irregular shapes. Moreover, these models are not generalizable across different tissues and markers.

Generative adversarial networks (GANs), introduced by Goodfellow et al. (21), have shown remarkable performance for a variety of image processing tasks including semantic segmentation of objects (22–24). Mahmood et al. (25) trained a supervised conditional GAN (cGAN) – that requires paired/co-registered training data – with synthetic and real data to overcome the multi-organ nuclei segmentation challenge. While the model showed promising results in segmenting nuclei, the performance degraded drastically on poor-quality input images or images where the assumed stain normalization failed or was not applied. We present a new stain-invariant multitask deep learning technique, DeepLIIF, which leverages cGAN and co-registered IHC and mpIF data to simultaneously translate IHC images to mpIF representations, accurately auto-segment relevant cells, and quantify protein expression for more accurate and reproducible IHC quantification. cGAN, with its combination of L1 loss and generator-discriminator framework, does away with the need for handcrafting loss functions for individual tasks (providing a seamless way for integrating additional tasks) and in contrast to the unsupervised counterparts, for example, CycleGAN, does not hallucinate or produce randomized outputs. Our model trained on clean IHC Ki67 images generalizes to more noisy and artifact-ridden images as well as other nuclear and non-nuclear markers such as CD3, CD8, BCL2, BCL6, MYC, MUM1, CD10 and TP53. As shown in Figure 1, given an IHC image, DeepLIIF simultaneously infers Hematoxylin (nuclear) channel, mpIF DAPI (nuclear), mpIF Lap2 (nuclear envelop), mpIF Ki67, and using these inferred modalities, automatically segments and classifies cells for accurate IHC quantification. Example IHC images stained with different markers along with the DeepLIIF inferred modalities and segmented/classified nuclear masks are also shown in Figure 1.

Results

In this section, we evaluate the performance of DeepLIIF on cell segmentation and classification tasks.

Metrics. We evaluated the performance of our model and other state-of-the-art methods using pixel accuracy (PixAcc) computed from the number of true positives, TP, false positives, FP and false negatives, FN, as $\frac{TP}{TP+FP+FN}$, Dice Score as $\frac{2\times TP}{2\times TP+FP+FN}$, and IOU as the class-wise intersection over the union. We compute these metrics for each class, including negative and positive, and compute the average value of both classes for each metric. A pixel is counted as TP if it is segmented and classified correctly. A pixel is considered FP if it is falsely segmented as the foreground of the corresponding class. A pixel is counted as FN if it is falsely detected as the background of the corresponding class. For example, assuming the model segments a pixel as a pixel of a negative cell (blue), but in the ground-truth mask, it is marked as positive (red). Since there is no corresponding pixel in the foreground of the ground-truth mask of the negative class, it is considered FP for the negative class and FN for the positive class, as there is no marked corresponding pixel in the foreground of the predicted mask of the positive class.

Testing Sets. To compare our model with state-of-the-art models, we use three different datasets. 1) We evaluate all models on our internal test set which includes 600 images of size 512×512 and 40x magnification from bladder carcinoma and non-small cell lung carcinoma slides. 2) We randomly selected and segmented 41 images of size 640×640 from recently released BCDataset (20) which contains Ki67 stained sections of breast carcinoma with Ki67+ and Ki67- cell centroid annotations (targeting cell detection as opposed to cell instance segmentation task). We split these tiles into 164 images of size 512×512; the test set varies widely in the density of tumor cells and the Ki67 index. 3) We also tested our model and others on a publicly available CD3 and CD8 IHC NuClick Dataset (26). We used the training set of this dataset containing 671 IHC patches of size 256×256, ex-

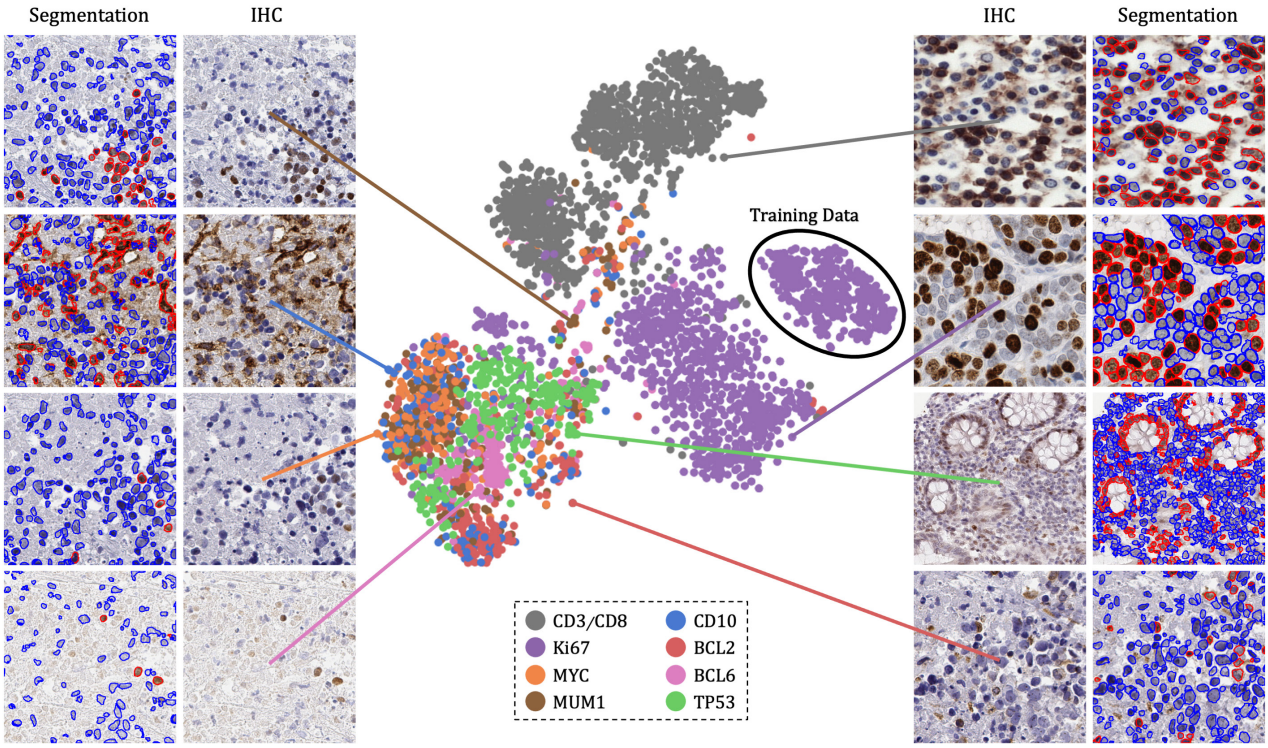


Fig. 2. The IHC markers in the tested datasets were embedded using t-SNE. Each point represents an IHC image of its corresponding marker. Randomly chosen example images of each marker are shown around the t-SNE plot. The black circle shows the cluster of training images.

tracted from LYON19 dataset (27). LYON19 is a challenge dataset for lymphocytes detection on IHC CD3/CD8 images taken from breast, colon and prostate.

Benchmarks. Trained on clean lung and bladder images stained with Ki67 marker, DeepLIIF generalizes well to other markers. We trained state-of-the-art segmentation networks, including *FPN* (28), *LinkNet* (29), *Mask_RCNN* (17), *Unet++* (16) on our training set (described in Sec A) using the IHC images as the input and generating the colored segmentation mask representing normal cells and lymphocytes. DeepLIIF outperformed previous models trained and tested on same data on all three metrics.

Evaluation. We compare the DeepLIIF model's performance against state-of-the-art models on the test set obtained from BCDataset (20). The results were analyzed both qualitatively and quantitatively, as shown in Figure 3. All models are trained and validated on the same training set as DeepLIIF model.

Application of DeepLIIF to the BC Dataset (20) resulted in a pixel accuracy of 84.00%, IOU of 36.70%, and Dice Score of 50.59%, and outperformed *Mask_RCNN* with pixel accuracy of 83.47%, IOU of 33.36%, and Dice Score of 46.64%, *UNet++* with pixel accuracy of 81.96%, IOU of 34.32%, and Dice Score of 47.88%, *LinkNet* with pixel accuracy of 82.59%, IOU of 35.05%, and Dice Score of 48.56%, and *FPN* with pixel accuracy of 80.64%, IOU of 34.25%, and Dice Score of 47.47%, while maintaining lower standard deviation on all metrics.

We used pixel-level accuracy metrics for the primary evaluation, as we are formulating the IHC quantification problem as cell instance segmentation/classification. However, since DeepLIIF is capable of separating the touching nuclei, we also performed a cell-level analysis of DeepLIIF against cell centroid detection approaches. *U−CSRNet* (20), for example, detects and classifies cells without performing cell instance segmentation. Most of these approaches use crowd counting techniques to find cell centroids. The major hurdle in evaluating these techniques is the variance in detected cell centroids. We trained *FCRN_A* (18), *FCRN_B* (18), *Deeplab_Xeption* (30), *SC_CNN* (19), *CSRNet* (31), *U−CSRNet* (20) using our training set (the centroids of our individual cell segmentation masks are used as detection masks). Most of these approaches failed in detecting and classifying cells on the BCData testing set, and the rest detected centroids far from the ground-truth centroids. As a result, we resorted to comparing the performance of DeepLIIF (trained on our training set) with these models trained on the training set of the BCData and tested on the testing set of the BCData. As shown in Extended Figure 1, even though our model was trained on a completely different dataset from the testing set, it has better performance than the detection models that were trained on the same training set of the test dataset. The results show that, unlike DeepLIIF, the detection models are not generalizable across different datasets, staining techniques, and tissue/cancer types.

As was mentioned earlier, our model generalizes well to segment/classify cells stained with different markers including CD3/CD8. We compare the performance of our trained model against other trained models on the training set of the NuClick dataset (32). The comparative analysis is shown

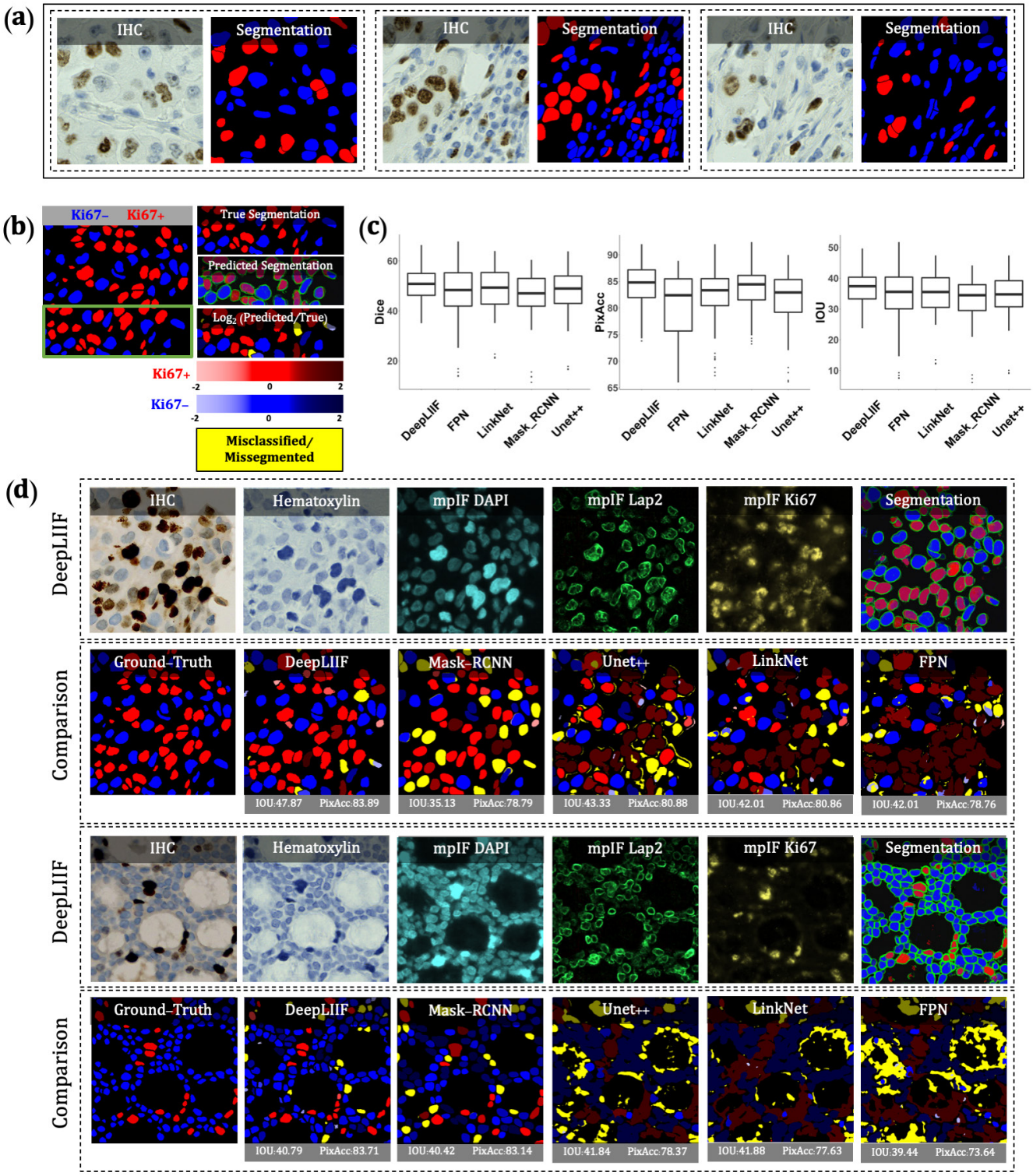


Fig. 3. (a) Three example images from our training set. (b) A segmentation mask showing Ki67- and Ki67+ cell representation, along with a visual segmentation and classification accuracy. Predicted classes are shown in different colors where blue represents Ki67- and red represents Ki67+ cells, and the hue is set using the log_2 of the ratio between the predicted area and ground-truth area. Cells with too large areas are shown in dark colors, and cells with too small areas are shown in a light color. For example, if the model correctly classifies a cell as Ki67+, but the predicted cell area is too large, the cell is colored in dark red. If there is no cell in the ground-truth mask corresponding to a predicted cell, the predicted cell is shown in yellow, which means that the cell is misclassified (cell segmented correctly but classified wrongly) or missegmented (no cell in the segmented cell area). (c) The accuracy of the segmentation and classification is measured by getting average of Dice score, Pixel Accuracy, and IOU (intersection over union) between the predicted segmentation mask of each class and the ground-truth mask of the corresponding class (0 indicates no agreement and 100 indicates perfect agreement). Evaluation of all scores show that DeepLIIF outperforms all state-of-the-art models. (d) As mentioned earlier, DeepLIIF generalizes across different tissue types and imaging platforms. Two example images from the BC Dataset (20) along with the generated modalities and classified segmentation masks are shown in the top rows where the ground-truth mask and segmentation masks of five state-of-the-art models are shown in the second row. The mean IOU and Pixel Accuracy are given for each generated mask.

in Figure 4. DeepLIIF model outperformed other models on segmenting and classifying CD3/CD8+ cells (tumor-infiltrating lymphocytes or TILs) on all three metrics.

We also tested DeepLIIF on other datasets including nine IHC snapshots from a digital microscope stained with Ki67 and PDL1 markers (two examples shown in Extended Data Figures 4 and 5), testing set of LYON19 (27) containing 441 ROIs (no annotations) from WSI of CD3/CD8 IHC specimens of breast, colon, and prostate (Figure 4(c), and Extended Data Figures 6, 7 and 8), Human Protein Atlas (33) IHC tiff images for TP53 (Figure 5), and the new DLBCL-Morph dataset (34) containing IHC tissue-microarrays for 209 patients stained with BCL2, BCL6, CD10, MYC, MUM1 markers (Figure 5 and Extended Data Figures 9, 10, 11, 12 and 13).

We have also evaluated the performance of DeepLIIF with and without LAP2beta and found the segmentation performance of DeepLIIF with LAP2beta better than without LAP2beta (Extended Data Figure 3). LAP2beta is a nuclear envelope protein broadly expressed in normal tissues. In Extended Data Figure 2, LAP2beta immunohistochemistry reveals nuclear envelope-specific staining in the majority of cells in spleen (99.98%), colon (99.41%), pancreas (99.50%), placenta (76.47%), testis (95.59%), skin (96.74%), lung (98.57%), liver (98.70%), kidney (95.92%) and lymph node (99.86%). Placenta syncytiotrophoblast does not stain with LAP2beta and the granular layer of skin does not show LAP2beta expression, however, the granular layer of skin lacks nuclei and is therefore not expected to express nuclear envelope proteins. We also observe lack of consistent Lap2beta staining in smooth muscle of blood vessel walls (not shown).

Discussion

Assessing IHC stained tissue sections is a widely utilized technique in diagnostic pathology laboratories worldwide. IHC-based protein detection in tissue with microscopic visualization is used for many purposes including tumor identification, tumor classification, cell enumeration as well as biomarker detection and quantification. Nearly all IHC stained slides for clinical care are analyzed and reported qualitatively or semi-quantitatively by diagnostic pathologists.

By creating a multitask deep learning framework referred to as DeepLIIF, we are providing a unified solution to nuclear segmentation and quantification of IHC stained slides. DeepLIIF is automated and does not require annotations. In contrast, most commercial platforms use a time-intensive workflow for IHC quantification which involves user-guided (a) IHC-DAB deconvolution, (b) nuclei segmentation of hematoxylin channel, (c) threshold setting for the brown DAB stain, and (d) cell classification based on the threshold. We present a simpler workflow; given an IHC input, we generate different modalities along with the segmented/classified cell masks. Our multitask deep learning framework performs IHC quantification in one process and does not require error-prone IHC deconvolution or manual thresholding steps. We use a single optimizer for all generators and discriminators that improves performance of all tasks simultaneously. Unique to this model, DeepLIIF is trained by generating registered mpIF, IHC and hematoxylin staining data from the same slide with the inclusion of nuclear envelope staining to assist in accurate segmentation of adjacent and overlapping nuclei.

Formulating the problem as cell instance segmentation/classification rather than a detection problem helps us to move beyond the reliance on crowd counting algorithms and towards more precise boundary delineation (semantic segmentation) and classification algorithms. DeepLIIF was trained for multi-organ, stain invariant determination of nuclear boundaries and classification of subsequent single cell nuclei as positive or negative for Ki67 staining detected with the 3,3'-Diaminobenzidine (DAB) chromogen. Subsequently, we determined that DeepLIIF accurately classified all tested nuclear antigens as positive or negative.

Surprisingly, DeepLIIF is often capable of accurate cell classification of non-nuclear staining patterns using CD3, CD8, BCL2, PDL1 and CD10. We believe the success of the DeepLIIF classification of non-nuclear markers is at least in part dependent on the the location of the chromogen deposition. BCL2 and CD10 protein staining often shows cytoplasmic chromogen deposition close to the nucleus and CD3 and CD8 most often stain small lymphocytes with scant cytoplasm whereby the chromogen deposition is physically close to the nucleus. DeepLIIF is slightly less accurate in classifying PDL1 staining (Extended Data Figure 5) and, notably, PDL1 staining is more often membranous staining of medium to large cells such as tumor cells and monocyte-derived cell lineages where DAB chromogen deposition is physically further from the nucleus. Since DeepLIIF was not trained for non-nuclear classification, we anticipate that further training using non-nuclear markers will rapidly improve their classification with DeepLIIF.

We have purposely assessed the performance of DeepLIIF for the detection of proteins currently reported semi-quantitatively by pathologists with the goal of facilitating transition to quantitative reporting. We anticipate further extension of this work to include validation of DeepLIIF on all markers in which more accurate, quantitative reporting would be clinically useful. Additional studies will also compare nuclear biomarker reporting for commonly used therapeutic targets such as ER, PR and AR. We will also assess the usability of Ki67 quantification in tumors with more unusual morphologic features such as sarcomas. The approach will also be extended to handle more challenging membranous/cytoplasmic markers such as PDL1, Her2, etc. Finally, we will incorporate additional mpIF tumor and immune markers into DeepLIIF for more precise phenotypic IHC quantification such as for distinguishing PDL1 expression within tumor vs. macrophage populations.

This work provides a universal, multitask model for both segmenting nuclei in IHC images and recognizing and quantifying positive and negative nuclear staining. Importantly, we describe a modality where training data from higher-cost and higher-dimensional multiplex imaging platforms improves

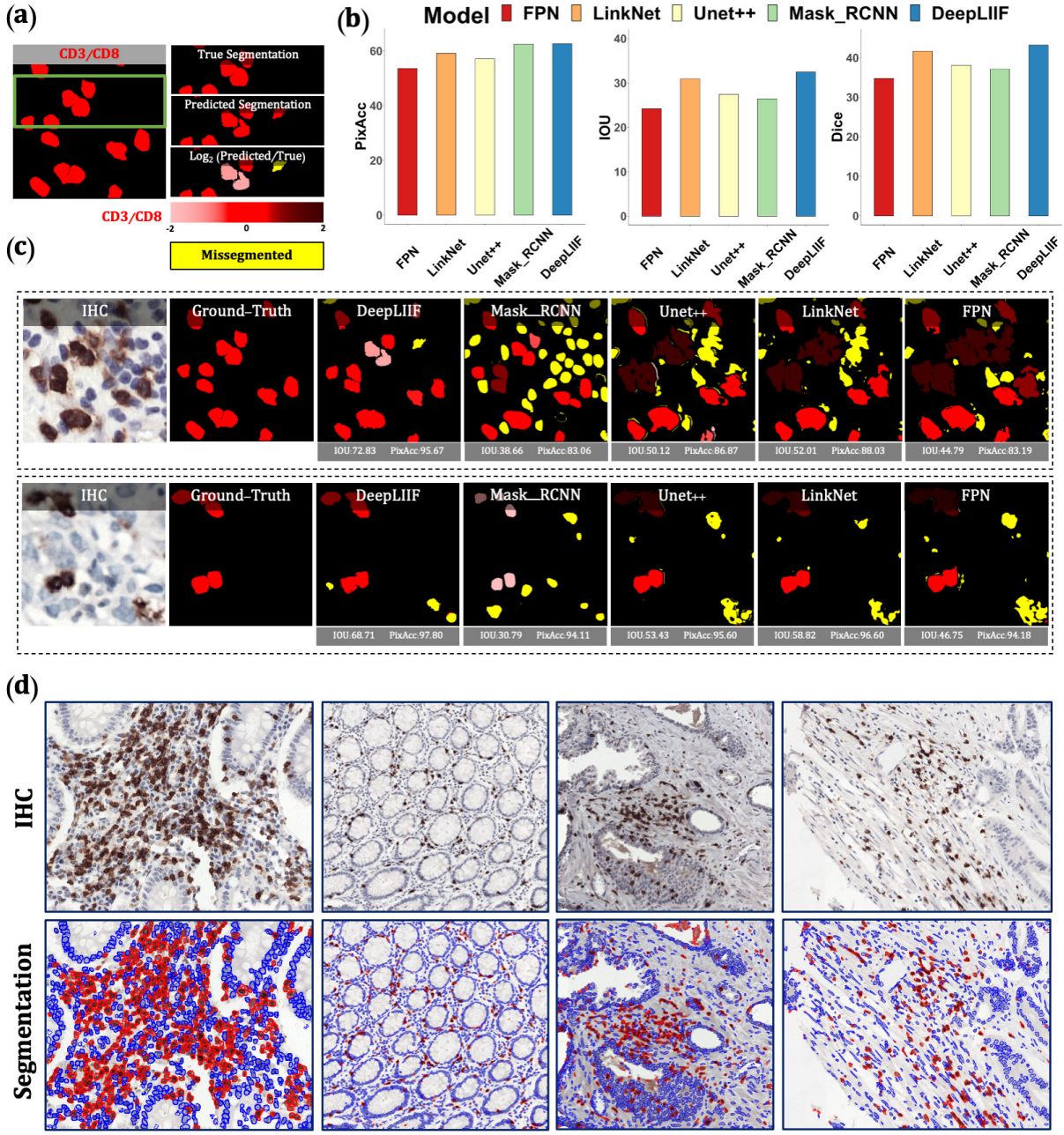


Fig. 4. (a) A segmentation mask showing CD3/CD8+ cells, along with a visual segmentation and classification accuracy. Predicted CD3/CD8+ cells are shown in red color, and the hue is set using the log_2 of the ratio between the predicted area and ground-truth area. Cells with too large areas are shown in dark colors, and cells with too small areas are shown in a light color. For example, if the model correctly classifies a cell as CD3/CD8+, but the predicted cell area is too large, the cell is colored in dark red. If there is no cell in the ground-truth mask corresponding to a predicted cell, the predicted cell is shown in yellow, which means that the cell is missegmented (no corresponding ground-truth cell in the segmented cell area). (b) The accuracy of the segmentation and classification is measured by getting average of Dice score, Pixel Accuracy, and IOU (intersection over union) between the predicted segmentation mask of CD3/CD8 and the ground-truth mask of the corresponding cells (0 indicates no agreement and 100 indicates perfect agreement). Evaluation of all scores show that DeepLIIF outperforms all state-of-the-art models. (c) As mentioned earlier, DeepLIIF generalizes across different tissue types and imaging platforms. Two example images from the NuClick Dataset (32) along with the modalities and classified segmentation masks generated by DeepLIIF are shown in the top rows where the ground-truth mask and quantitative segmentation masks of DeepLIIF and state-of-the-art models are shown in the second row. The mean IOU and Pixel Accuracy are given for each generated mask. (d) Randomly chosen samples from the LYON19 challenge dataset (27). The top row shows the IHC image and the bottom row shows the classified segmentation mask generated by DeepLIIF. In the mask, the blue color shows the boundary of negative cells and the red color shows the boundary of positive cells.

the interpretability of more widely-used and lower-cost IHC.

Methods

A. Training Data. To train DeepLIIF, we used a dataset of lung and bladder tissues containing IHC, hematoxylin, mpIF DAPI, mpIF Lap2, and mpIF Ki67 of the same tissue scanned using ZEISS Axioscan. These images were scaled and co-registered with the fixed IHC images using affine transformations, resulting in 1667 registered sets of IHC images and the other modalities of size 512×512. We randomly selected 709 sets for training, 358 sets for validation, and 600 sets for testing the model.

Ground-truth Classified Segmentation Mask. To create the ground-truth segmentation mask for training and testing our model, we used our interactive deep learning ImPartial annotations framework (35). Given mpIF DAPI images and few cell annotations, this framework auto-thresholds and performs cell instance segmentation for the entire image. Using this framework, we generated nuclear segmentation masks for each registered set of images with precise cell boundary delineation. Finally, using the mpIF Ki67 images in each set, we classified the segmented cells in the segmentation mask, resulting in 9180 Ki67 positive cells and 59000 Ki67 negative cells. Examples of classified segmentation masks from the ImPartial framework are shown in Figures 1 and 3. The white boundary around the cells are generated by ImPartial, and the cells are classified into red (positive) and blue (negative) using the corresponding mpIF Ki67 image. If a segmented cell has any representation in the mpIF Ki67 image, we classify it as positive (red color), otherwise, we classify it as negative (blue color).

B. Objective. Given a dataset of IHC+Ki67 RGB images, our objective is to train a model $f(.)$ that maps an input image to four individual modalities, including Hematoxylin channel, mpIF DAPI, mpIF Lap2, and mpIF Ki67 images, and using the mapped representations, generate the segmentation mask. We present a framework, as shown in Figure 6 that performs two tasks simultaneously. First, the translation task translates the IHC+Ki67 image into four different modalities for clinical interpretability as well as for segmentation. Second, a segmentation task generates a single classified segmentation mask from the IHC input and three of the inferred modalities by applying a weighted average and coloring cell boundaries green, positive cells red, and negative cells blue.

We use cGANs to generate the modalities and the segmentation mask. cGANs are made of two distinct components, a generator and a discriminator. The generator learns a mapping from the input image x to output image y, $G : x \rightarrow y$. The discriminator learns to the paired input and output of the generator from the paired input and ground truth result. We define eight generators to produce four modalities and segmentation masks that cannot be distinguished from real images by eight adversarially trained discriminators (trained to detect fake images from the generators).

Translation. Generators G_{t_1}, G_{t_2}, G_{t_3}, and G_{t_4} produce hematoxylin, mpIF DAPI, mpIF Lap2, and mpIF Ki67 images from the input IHC image, respectively ($G_{t_i} : x_i \rightarrow y_i$, where $i = 1, 2, 3, 4$). The discriminator D_i is responsible for discriminating generated images by generators G_{t_i}. The objective of the conditional GAN for the image translator tasks are defines as follows:

$$\begin{aligned}\mathcal{L}_{tGAN}(G_{t_i}, D_{t_i}) = \mathbb{E}_{x,y_i}[\log D_{t_i}(x, y_i)] \\ + \mathbb{E}_{x,y_i}[\log(1 - D_{t_i}(x, G_{t_i}(x)))]\end{aligned} \quad \textbf{(1)}$$

We use smooth L1 loss (Huber loss) to compute the error between the predicted value and the true value, since it is less sensitive to outliers compared to L2 loss and prevents exploding gradients while minimizing blur (36, 37). It is defined as:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[smooth_{L1}(y - G(x))] \quad \textbf{(2)}$$

where

$$smooth_{L1}(a) = \begin{cases} 0.5a^2 & \text{if } |a| < 0.5 \\ |a| - 0.5 & \text{otherwise} \end{cases} \quad \textbf{(3)}$$

The objective loss function of the translation task is:

$$\mathcal{L}_T(G_t, D_t) = \sum_{i=1 \sim 5} \mathcal{L}_{tGAN}(G_{t_i}, D_{t_i}) + \mathcal{L}_{L1}(G_{t_i}) \quad \textbf{(4)}$$

Segmentation/Classification. The segmentation component consists of four generators G_{S_1}, G_{S_2}, G_{S_3}, and G_{S_4} producing four individual segmentation masks from the real IHC Ki67 image, generated hematoxylin image (G_{t_1}), generated mpIF DAPI (G_{t_2}), and generated mpIF Lap2 (G_{t_3}), $G_{S_i} =: z_i \rightarrow y_{s_i}$ where $i = 1, 2, 3, 4$. The final segmentation mask of G_S generator is created by averaging the four generated segmentation masks by G_{S_i} using pre-defined weights, $S(z_i) = \sum_{n=1}^{4} ws_i \times G_{S_i}(z_i)$, where ws_i are the pre-defined weights. The discriminators D_{S_i} are responsible for discriminating generated images by generators G_{S_i}.

In this task, we use LSGAN loss function, since it solves the problem of vanishing gradients for the segmented pixels on the correct side of the decision boundary, but far from the real data, resulting in a more stable boundary segmentatin learning process. We define the objective of the conditional GAN for segmentation/classification task as follows:

$$\begin{aligned}\mathcal{L}_{sGAN}(D_S) = \sum_{i=1 \sim 4} (\frac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, y_{s_i}) - 1)^2] \\ + \frac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, S(z_i)))^2]) \\ \mathcal{L}_{sGAN}(S) = \sum_{i=1 \sim 4} \frac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, S(z_i)) - 1)^2]\end{aligned} \quad \textbf{(5)}$$

For this task, we also use smooth L1 loss. The objective loss function of the segmentation/classification task is:

$$\mathcal{L}_S(S, D_S) = \mathcal{L}_{sGAN}(S, D_S) + \mathcal{L}_{L1}(S) \quad \textbf{(6)}$$

Final Objective. The final objective is:

$$\begin{aligned}\mathcal{L}(G_t, D_t, S, D_S) = \mathcal{L}_T(G_t, D_t) \\ + \mathcal{L}_S(S, D_S)\end{aligned} \quad \textbf{(7)}$$

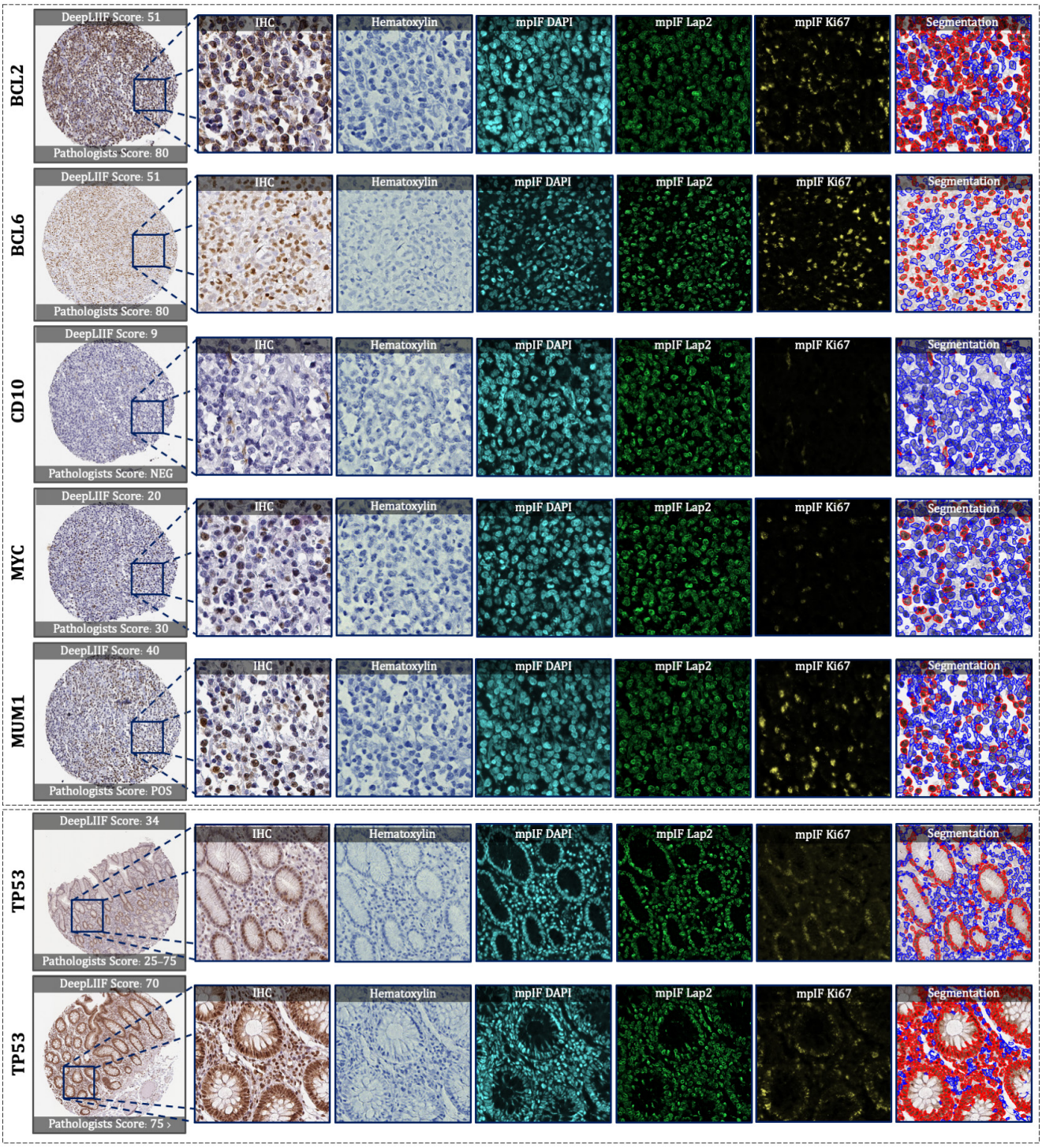


Fig. 5. Examples of tissues stained with various markers. The top box shows sample tissues stained with BCL2, BCL6, CD10, MYC, and MUM1 from DLBCL-morph dataset (34). The bottom box shows sample images stained with TP53 marker from the Human Protein Atlas (33). In each row, the first image on the left shows the original tissue stained with a specific marker. The quantification score computed by the classified segmentation mask generated by DeepLIIF is shown on the top of the whole tissue image, and the predicted score by pathologists is shown on the bottom. In the following images of each row, the modalities and the classified segmentation mask of a chosen crop from the original tissue are shown.

C. Generator. We use two different types of generators, ResNet-9blocks generator for producing modalities and U-Net generator for creating segmentation mask.

C.1. ResNet-9blocks Generator. The generators responsible for generating modalities including hematoxylin, mpIF DAPI and mpIF Lap2 starts with a convolution layer and a batch normalization layer followed by Rectified Linear Unit (ReLU) activation function, 2 downsampling layers, 9 residual blocks, 2 upsampling layers, and a covolutional layer followed by a tanh activation function. Each residual block consists of two convolutional layers with the same number

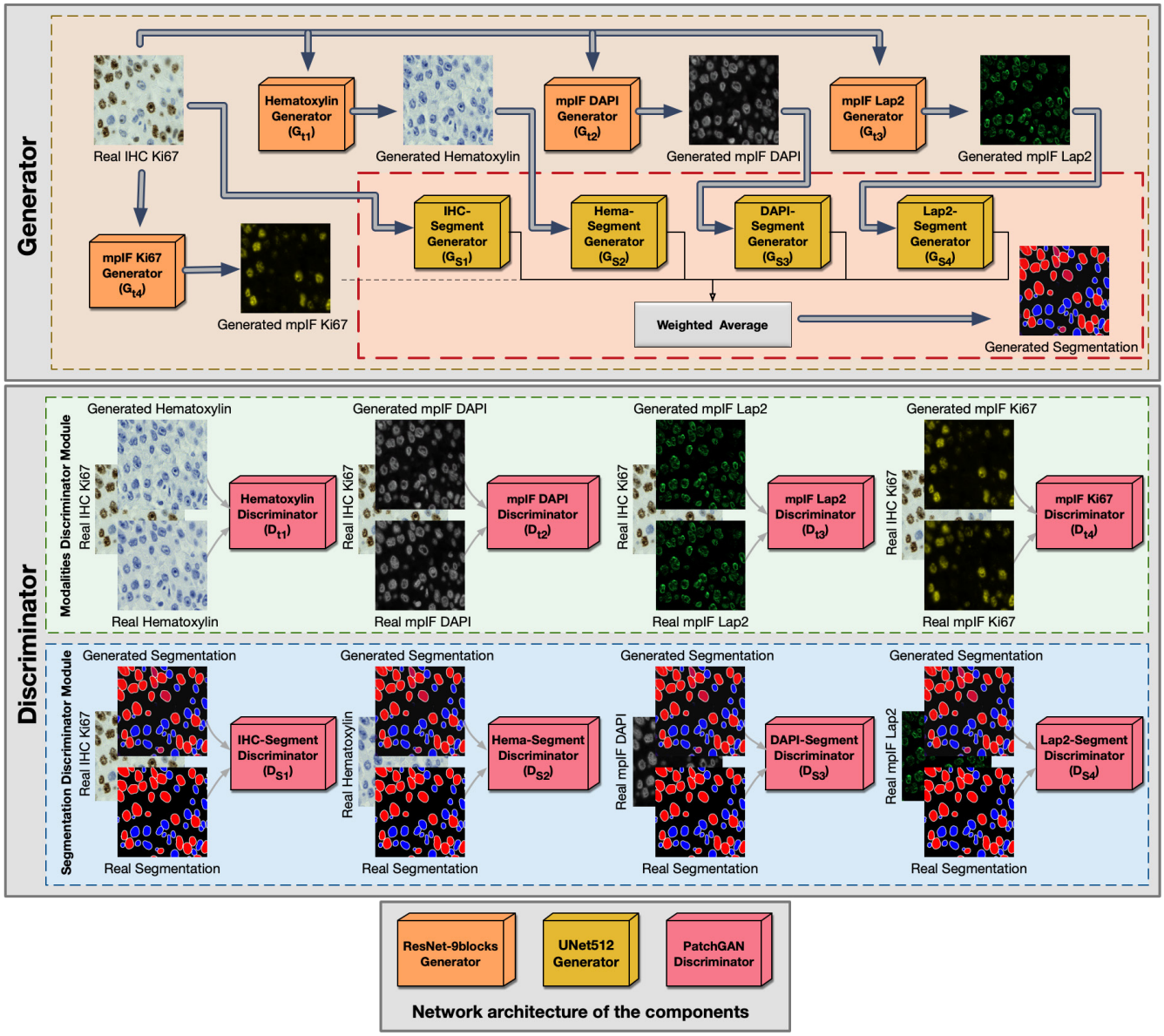


Fig. 6. Overview of DeepLIIF. The network consists of a generator and a discriminator component. It uses ResNet-9block generator for generating the modalities including Hematoxylin, mpIF DAPI, mpIF Lap2, and mpIF Ki67 and UNet512 generator for generating the segmentation mask. In the segmentation component, the generated masks from IHC, Hematoxylin, mpIF DAPI, and mpIF Lap2 representations are averaged with pre-defined weights to create the final segmentation mask. The discriminator component consists of the modalities discriminator module and segmentation discriminator module.

of output channels. Each convolutional layer in the residual block is followed by a batch normalization layer and a ReLU activation function. Then, these convolution operations are skipped and the input is directly added before the final ReLU activation function.

C.2. U-Net Generator. For generating the segmentation masks, we use the generator proposed by (37), using the general shape of U-Net (38) with skip connections. The skip connections are added between each layer i and layer $n-i$ where n is the total number of layers. Each skip connection concatenates all channels at layer i with those at layer $n-i$.

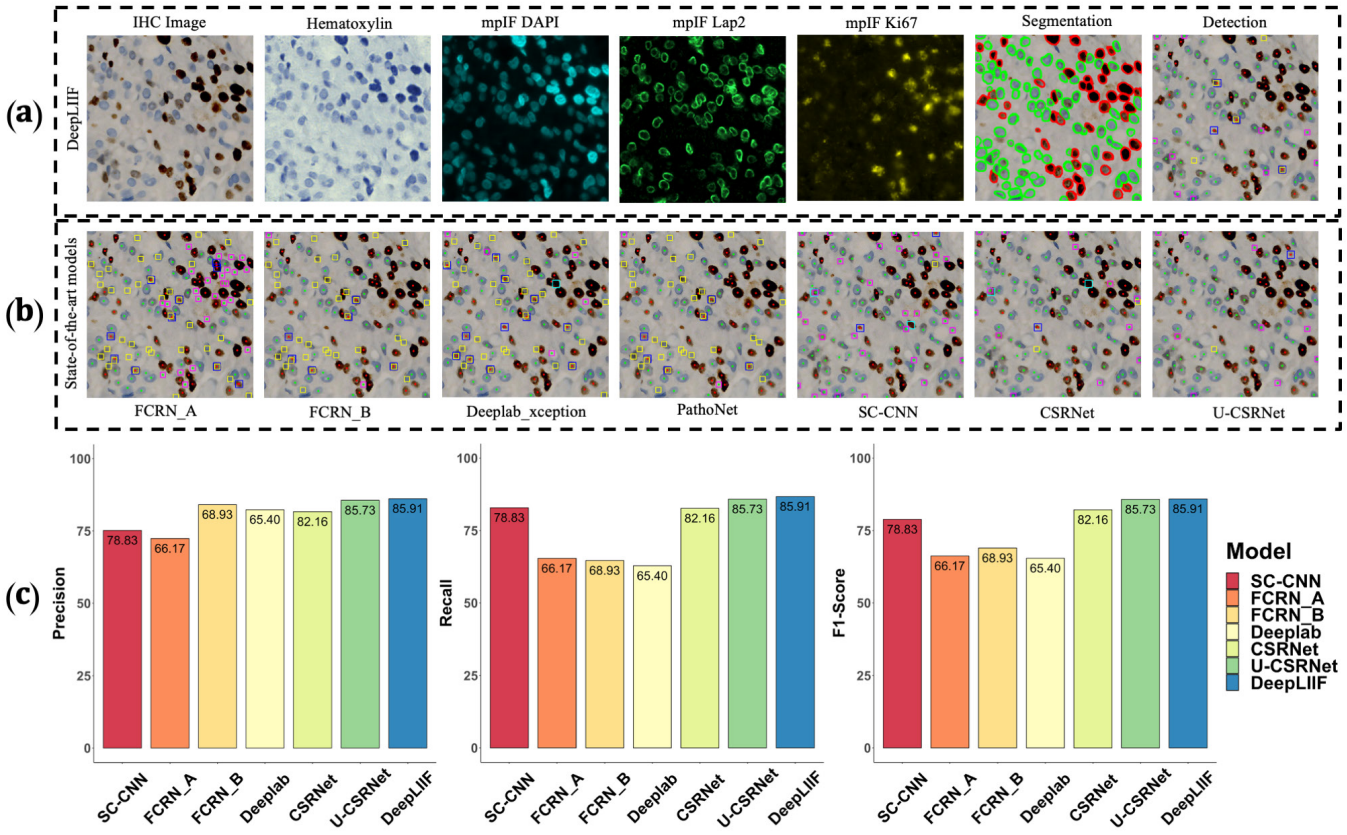
D. Markovian discriminator (PatchGAN). To address high-frequencies in the image, we use a PatchGAN discriminator that only penalizes structure at the scale of patches. It classifies each $N \times N$ patch in an image as real or fake. We run this fully convolutional discriminator across the image, averaging all responses to provide the final output of D.

E. Optimization. To optimize our network, we use the same standard approach as (21), alternating between one gradient descent step on D and one step on G. In all defined tasks (translation, classification, and segmentation), the network generates different representations for the same cells in the input meaning all tasks have the same endpoint. Therefore, we use a single optimizer for all generators and a single optimizer for all discriminators. Using this approach, optimizing the parameters of a task with a more clear representation of cells improves the accuracy of other tasks since all these tasks are optimized simultaneously.

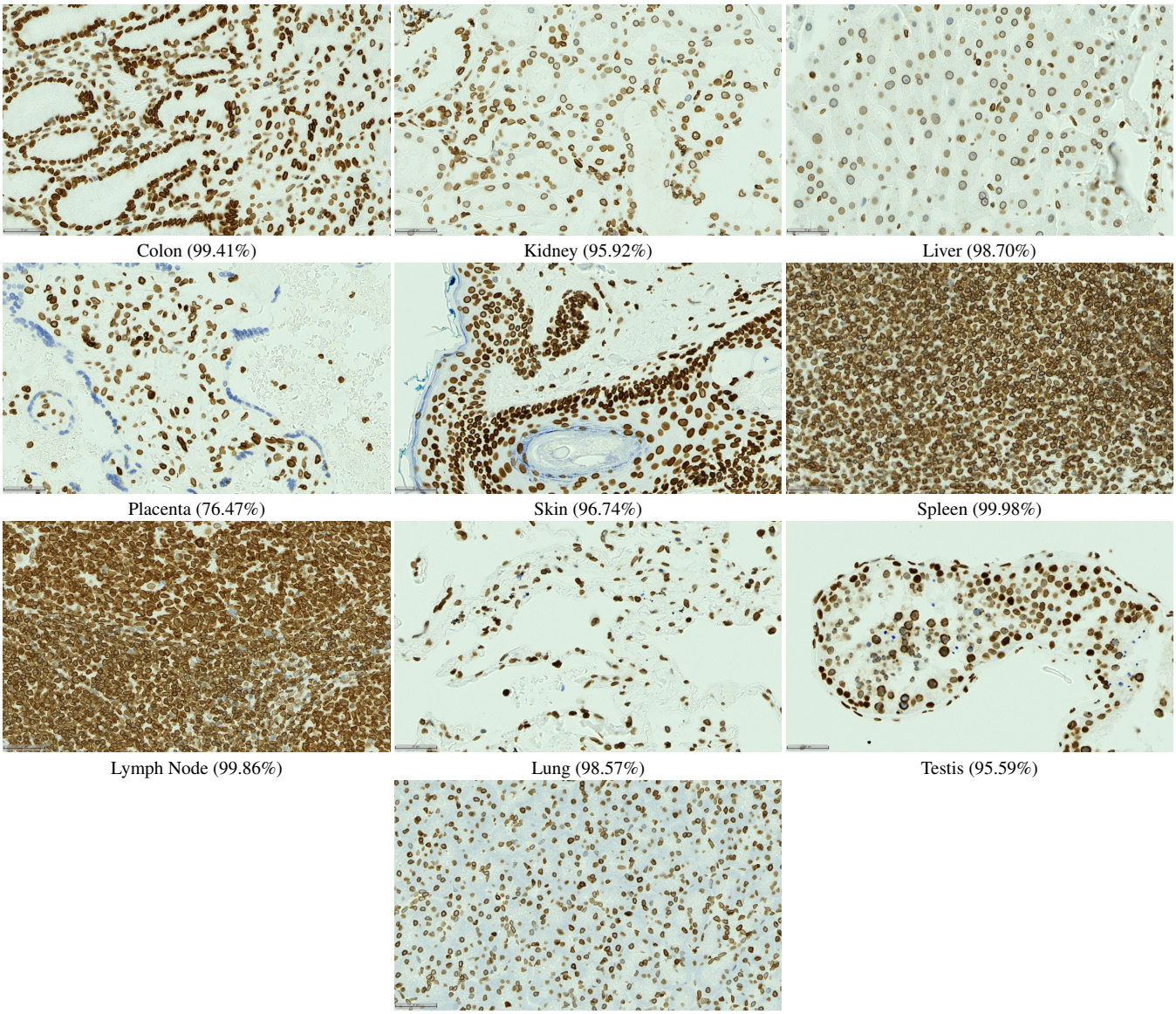
ACKNOWLEDGEMENTS
This project was supported by MSK Cancer Center Support Grant/Core Grant (P30 CA008748) and partly by MSK DigITs Hybrid Research Initiative, and in part by NSF grants CNS1650499, OAC1919752, and ICER1940302.

Bibliography

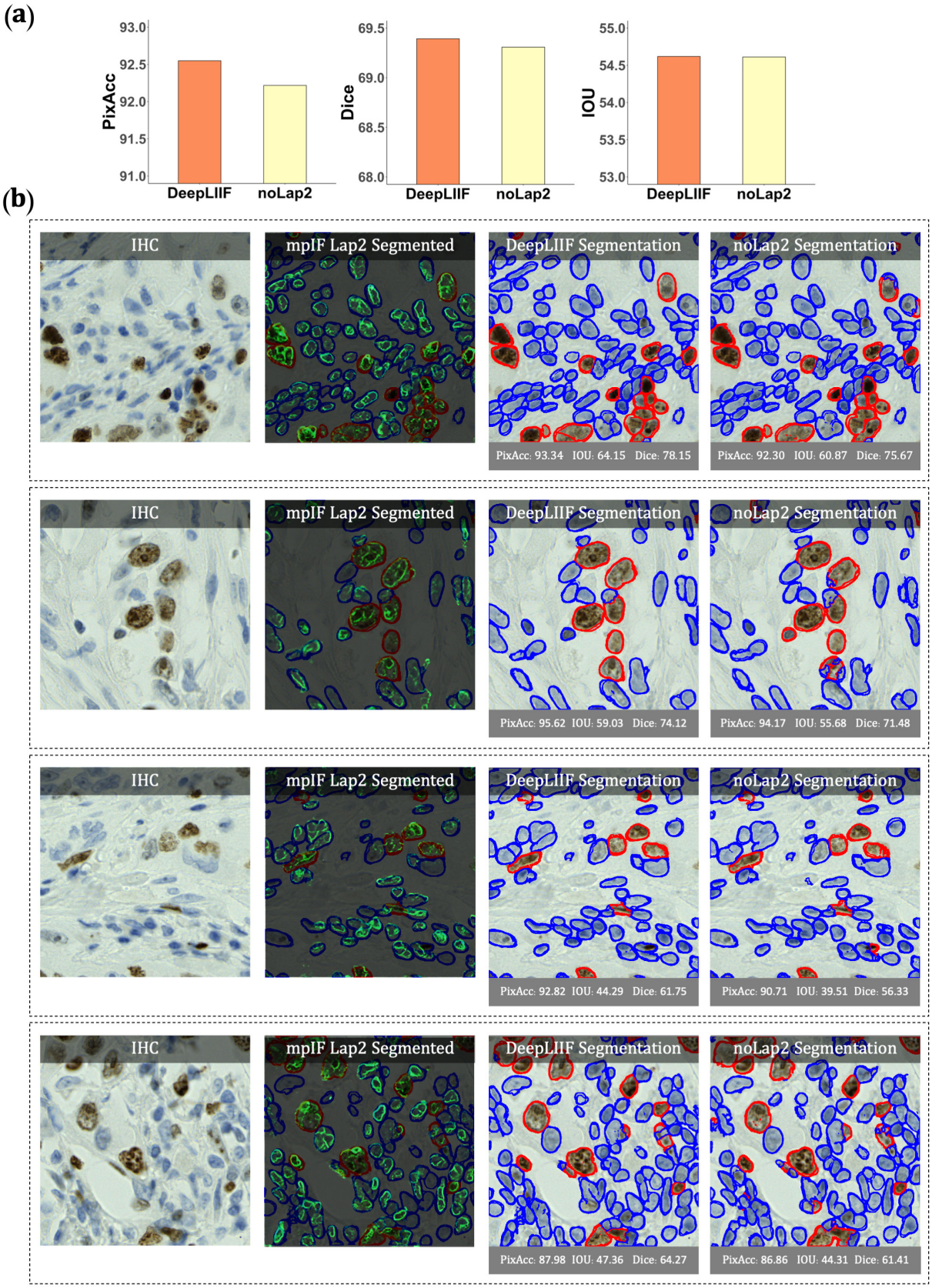
1. Wei Chang Colin Tan, Sanjna Nilesh Nerurkar, Hai Yun Cai, Harry Ho Man Ng, Duoduo Wu, Yu Ting Felicia Wee, Jeffrey Chun Tatt Lim, Joe Yeong, and Tony Kiat Hon Lim. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, 40(4):135–153, 2020.
2. Joe Yeong, Tira Tan, Zi Long Chow, Qing Cheng, Bernett Lee, Amanda Seet, Johnathan Xiande Lim, Jeffrey Chun Tatt Lim, Clara Chong Hui Ong, Aye Aye Thike, et al. Multiplex immunohistochemistry/immunofluorescence (mihc/if) for pd-l1 testing in triple-negative breast cancer: a translational assay compared with conventional ihc. *Journal of Clinical Pathology*, 2020.
3. Steve Lu, Julie E Stein, David L Rimm, Daphne W Wang, J Michael Bell, Douglas B Johnson, Jeffrey A Sosman, Kurt A Schalper, Robert A Anders, Hao Wang, et al. Comparison of biomarker modalities for predicting response to pd-1/pd-l1 checkpoint blockade: a systematic review and meta-analysis. *JAMA oncology*, 5(8):1195–1204, 2019.
4. Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
5. Maximilian Ernst Tschuchnig, Gertie Janneke Oostingh, and Michael Gadermayr. Generative adversarial networks in digital pathology: A survey on trends and future potential. *Pattern*, 1(6):100089, 2020.
6. Yair Rivenson, Kevin de Haan, W Dean Wallace, and Aydogan Ozcan. Emerging advances to transform histopathology using virtual staining. *BME Frontiers*, 2020, 2020.
7. 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. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4243–4252, 2020.
8. Erik A Burlingame, Mary McDonnell, Geoffrey F Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E Johnson, Christopher Corless, Joe W Gray, and Young Hwan Chang. Shift: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Scientific reports*, 10(1):1–14, 2020.
9. Caner Mercan, GCAM Mooij, David Tellez, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1770–1774, 2020.
10. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.
11. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.
12. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. doi: 10.1109/TPAMI.2017.2699184.
13. Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.
14. Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. doi: 10.1109/TPAMI.2019.2893630.
15. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. doi: https://doi.org/10.1007/978-3-319-24574-4_28.
16. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.
17. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
18. Weidi Xie, J. Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):283–292, 2018. doi: 10.1080/21681163.2016.1149104.
19. S. Ram and J. J. Rodríguez. Size-invariant detection of cell nuclei in microscopy images. *IEEE Transactions on Medical Imaging*, 35(7):1753–1764, 2016. doi: 10.1109/TMI.2016.2527740.
20. Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Ruizhe Geng, Hongliang He, Shan Du, Xia Liu, Yonghong Tian, Yongsheng Liang, S. Kevin Zhou, and Jie Chen. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 289–298. Springer International Publishing, 2020. ISBN 978-3-030-59722-1.
21. 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.
22. Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. doi: 10.1109/ICCV.2017.606.
23. Xinming Zhang, Xiaobin Zhu, Naiguang Zhang, Peng Li, Lei Wang, et al. Seggan: Semantic segmentation with generative adversarial network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5. IEEE, 2018. doi: 10.1109/BigMM.2018.8499105.
24. Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 6830–6840, 2019. doi: 10.1109/ICCV.2019.00693.
25. 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. doi: 10.1109/TMI.2019.2927182.
26. Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.
27. Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijthoven, Maschenka Balkenhol, Margarita Melnikova, Oscar Geessink, Quirine Manson, Mark Sherman, Antonio Polonia, Jeremy Parry, et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Medical image analysis*, 58:101547, 2019.
28. Alexander Kirillov, Kaiming He, Ross Girshick, and Piotr Dollár. A unified architecture for instance and semantic segmentation, 2017.
29. Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.
30. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
31. Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.
32. Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101771.
33. Andreas Digre and Cecilia Lindskog. The human protein atlas—spatial localization of the human proteome in health and disease. *Protein Science*, 30(1):218–233, 2021.
34. Damir Vrabac, Akshay Smit, Rebecca Rojansky, Yasodha Natkunam, Ranjana H Advani, Andrew Y Ng, Sebastian Fernandez-Pol, and Pranav Rajpurkar. Dlbcl-morph: Morphological features computed using deep learning for an annotated digital dlbcl image set. *arXiv preprint arXiv:2009.08123*, 2020.
35. Natalia Martinez, Guillermo Sapiro, Allen Tannenbaum, Travis J. Hollmann, and Saad Nadeem. Impartial: Partial annotations for cell instance segmentation. *bioRxiv*, 2021. doi: 10.1101/2021.01.20.427458.
36. Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
37. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.
38. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015.



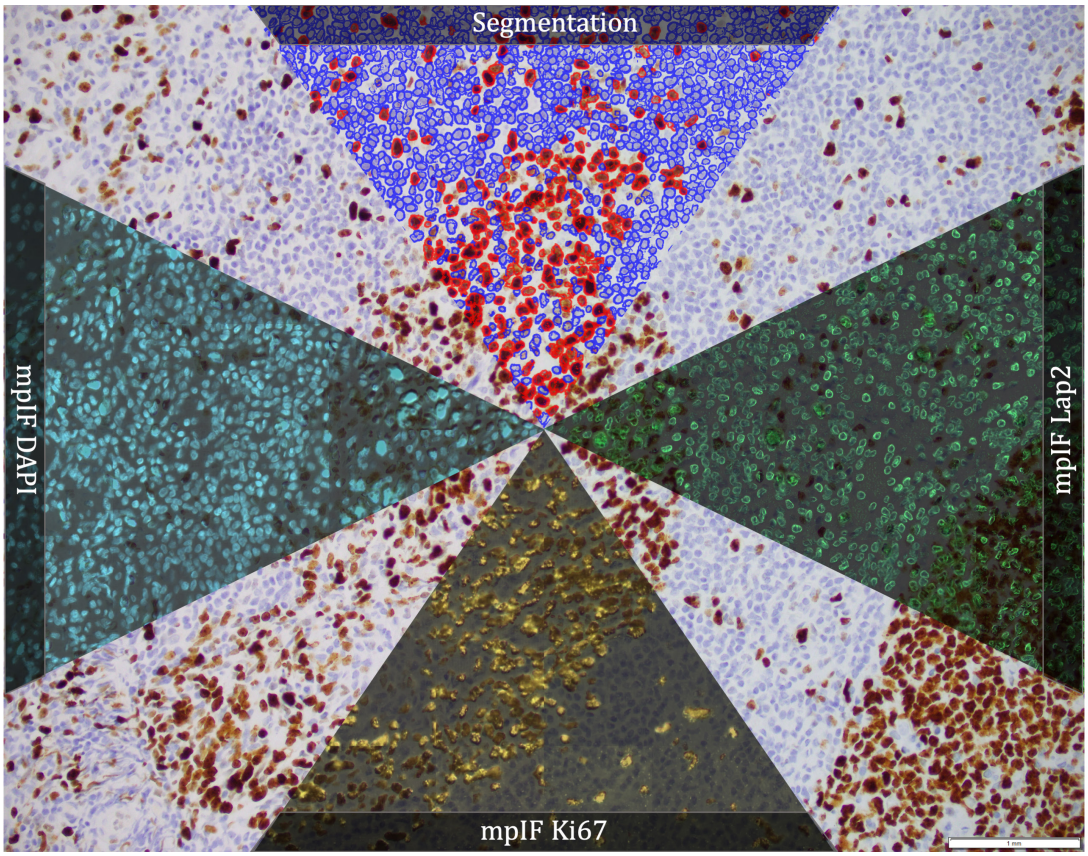
Extended Data Fig. 1. Qualitative and quantitative analysis of DeepLIIF against detection models on the testing set of the BC Data (20). (a) An example IHC image from the BC Data testing set, the generated modalities, segmentation mask overlaid on the IHC image, and the detection mask generated by DeepLIIF. (b) The detection masks generated by the detection models. In the detection mask, the center of a detected positive cell is shown with red dot and the center of a detected negative cell is shown with blue dot. We show the missing positive cells in cyan bounding boxes, the missing negative cells in yellow bounding boxes, the wrongly detected positive cells in blue bounding boxes, the wrongly detected negative cells in pink bounding boxes. (c) The detection accuracy is measured by getting average of precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), and f1-score ($\frac{2\times precision\times recall}{precision+recall}$) between the predicted detection mask of each class and the ground-truth mask of the corresponding class. A predicted point is regarded as true positive if it is within the region of a ground-truth point with a predefined radius (we set it to 10 pixels in our experiment which is similar to the predefined radius in (20)). Centers that have been detected more than once are considered as false positive. Evaluation of all scores show that DeepLIIF outperforms all state-of-the-art models.



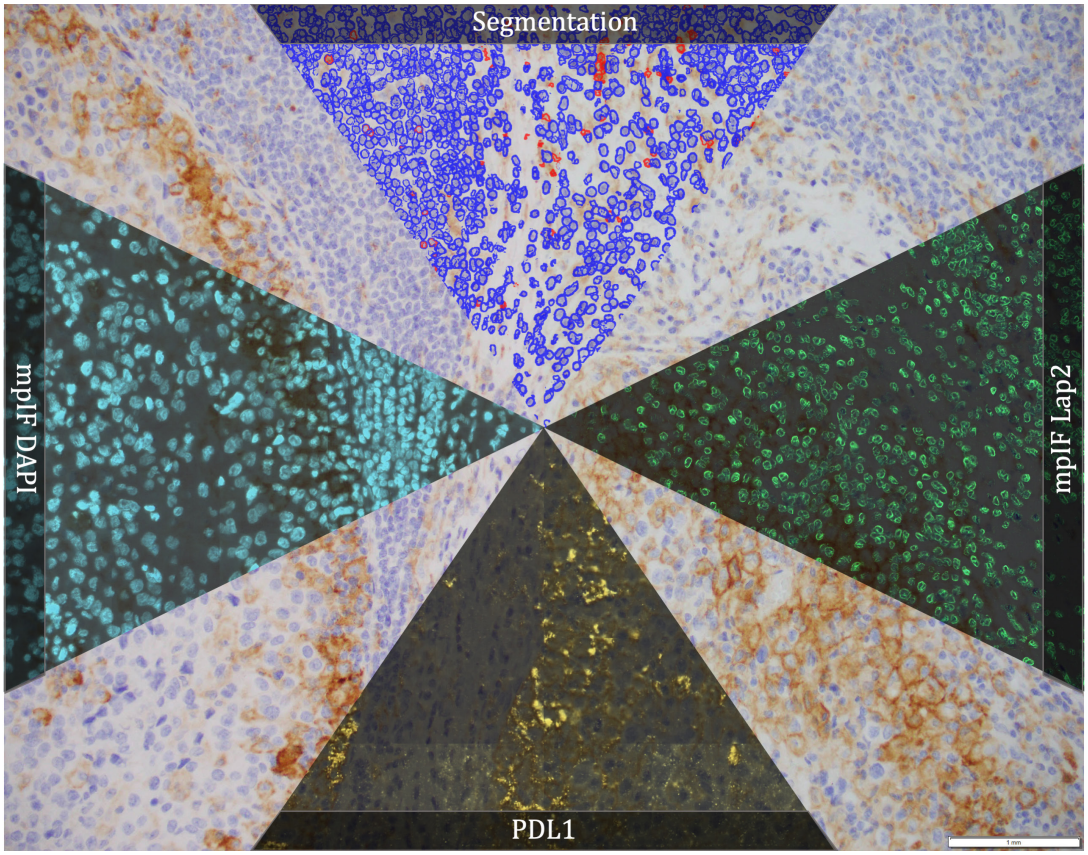
Extended Data Fig. 2. LAP2beta coverage for normal tissues. LAP2beta immunohistochemistry reveals nuclear envelope-specific staining in the majority of cells in spleen (99.98%), colon (99.41%), pancreas (99.50%), placenta (76.47%), testis (95.59%), skin (96.74%), lung (98.57%), liver (98.70%), kidney (95.92%) and lymph node (99.86%).



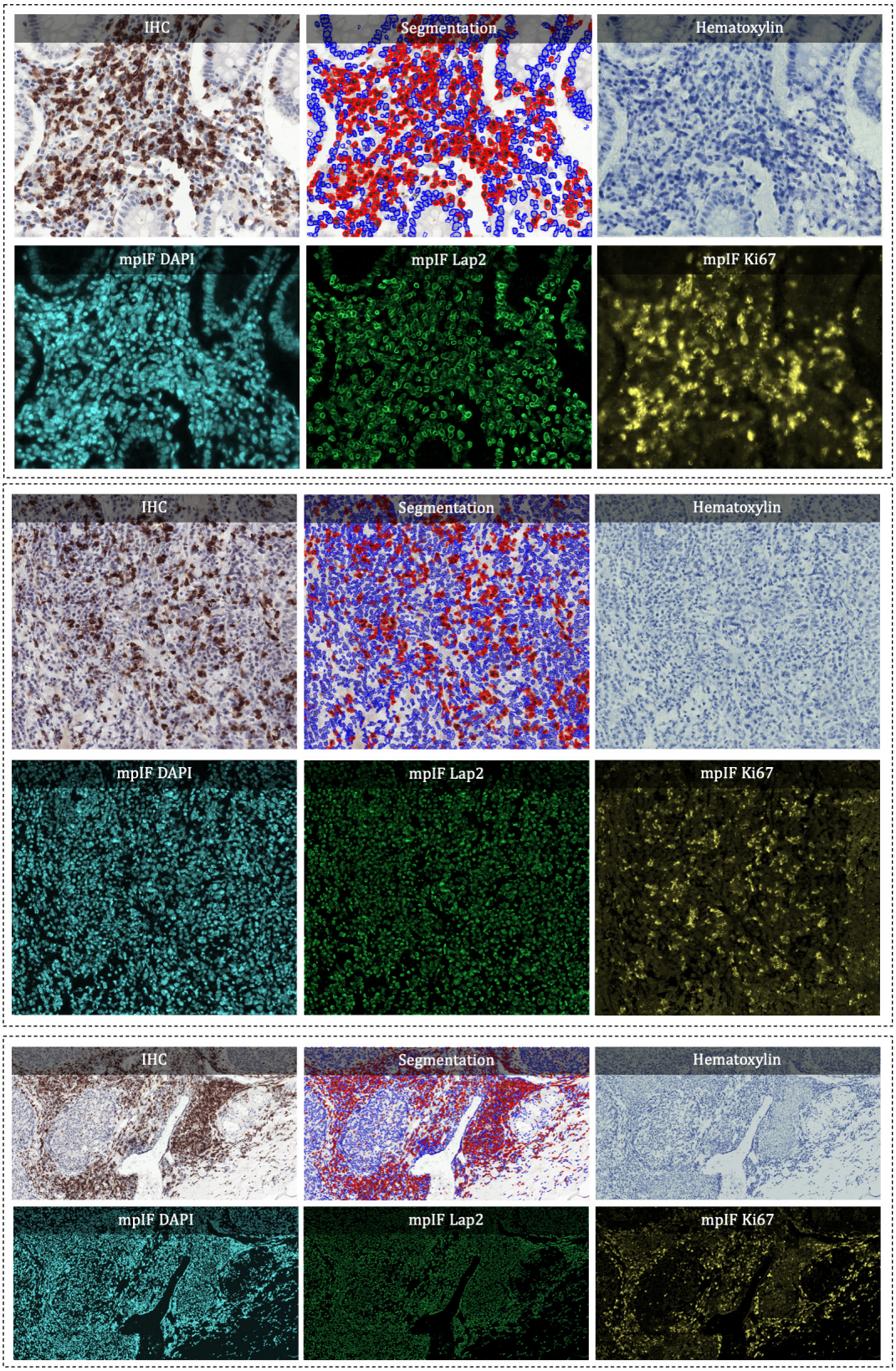
Extended Data Fig. 3. (a) A qualitative comparison of DeepLIIF against noLap2 model. (b) Some example IHC images. The first image in each row shows the input IHC image. In the second image, the generated mpIF Lap2 image is overlaid on the classified/segmented IHC image. The third and fourth images show the segmentation mask, respectively, generated by DeepLIIF and noLap2.



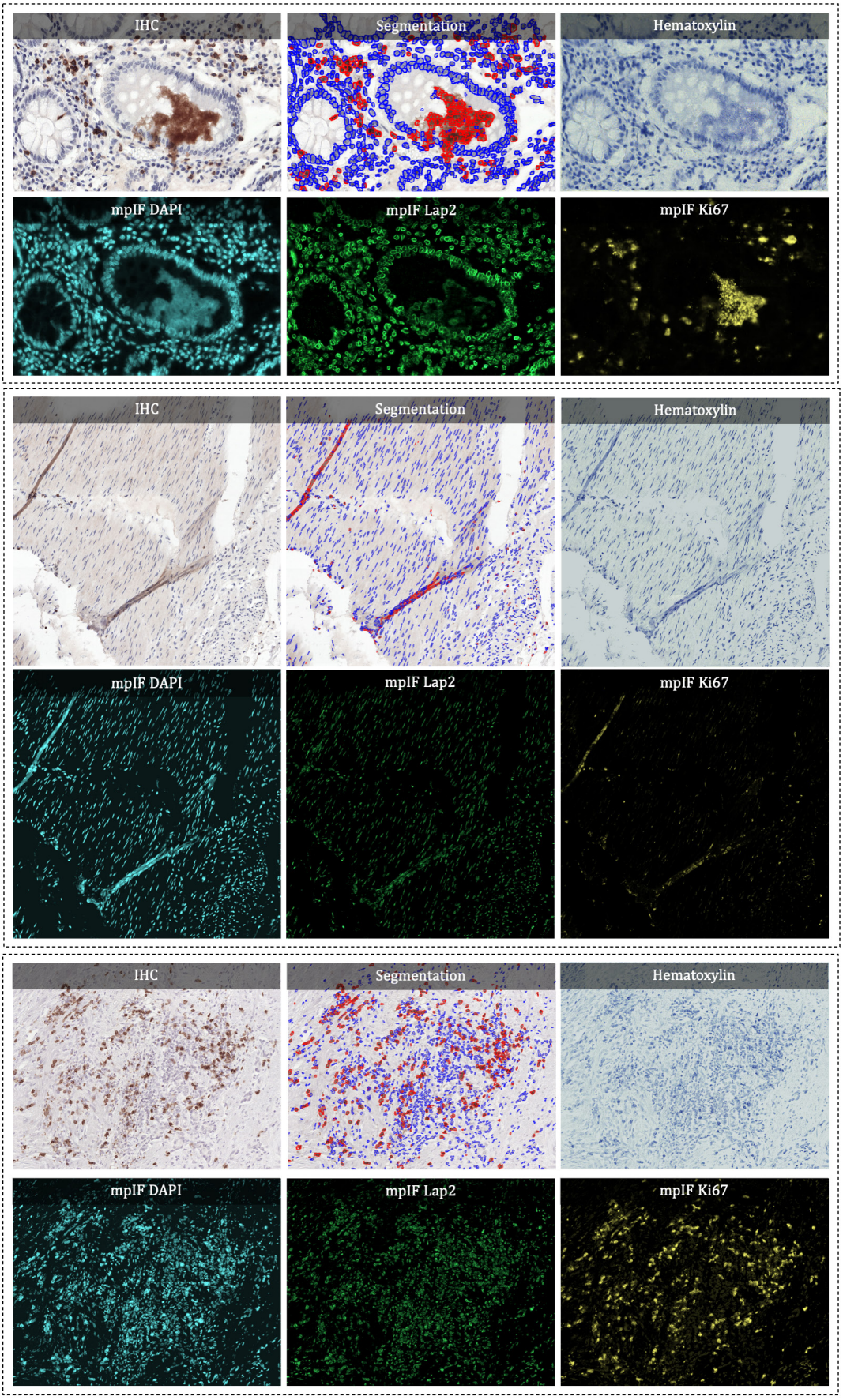
Extended Data Fig. 4. Microscope Snapshot for IHC Ki67.



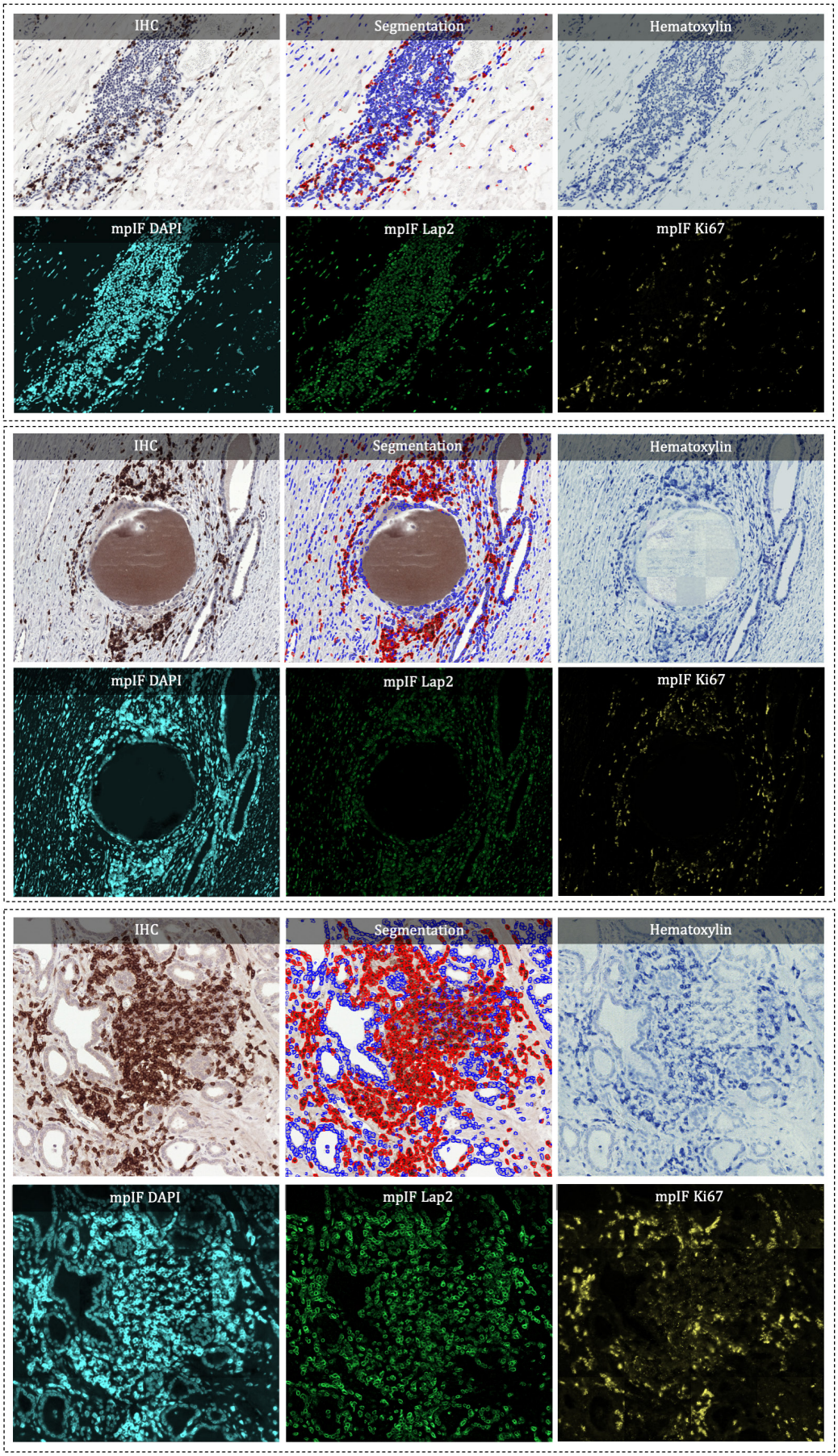
Extended Data Fig. 5. Microscopic Snapshot for IHC PDL1.



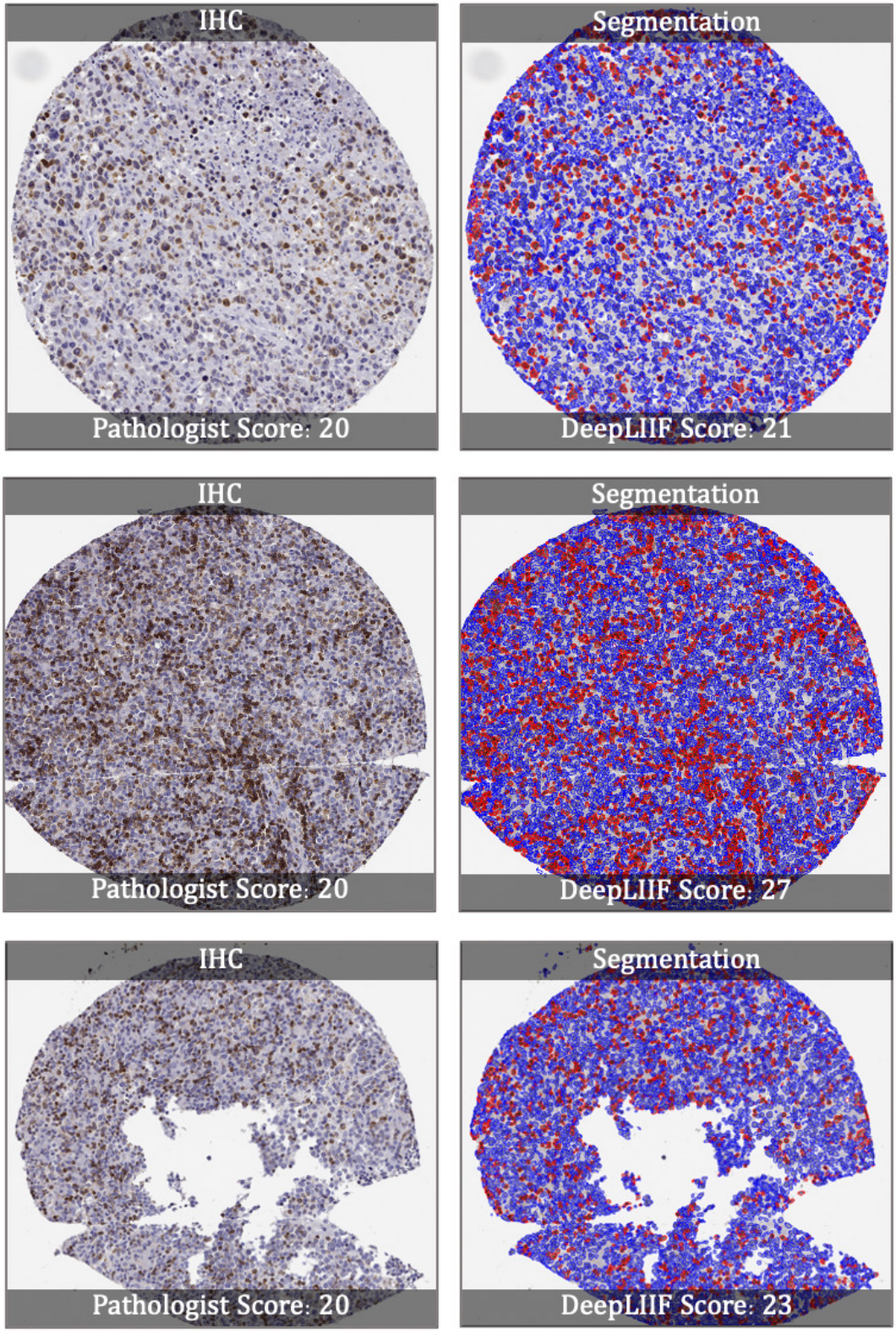
Extended Data Fig. 6. Some examples from LYON19 Challenge Dataset (27). The generated modalities and classified segmentation mask for each sample are shown in a separate box.



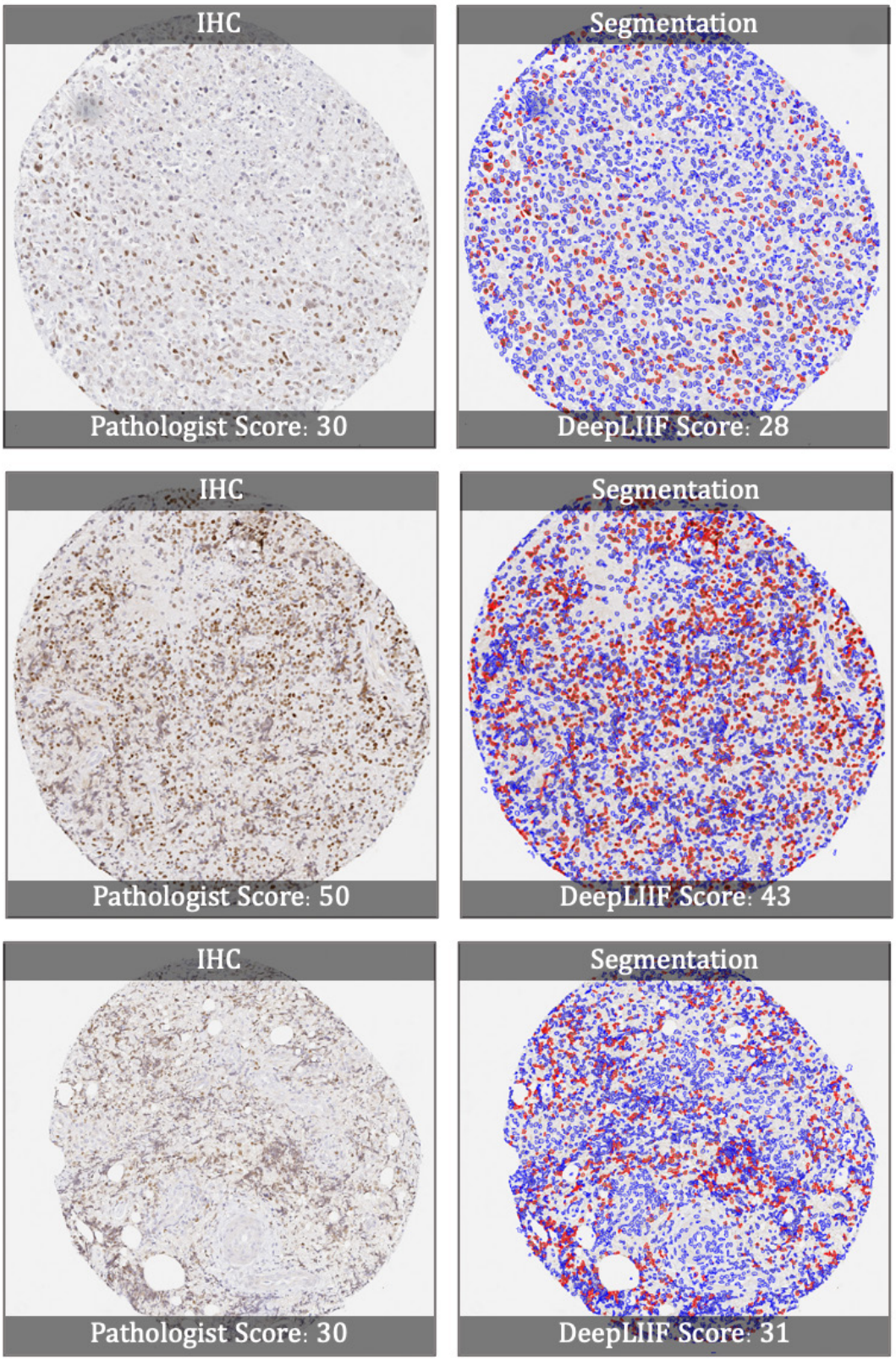
Extended Data Fig. 7. Some examples from LYON19 Challenge Dataset (27). The generated modalities and classified segmentation mask for each sample are shown in a separate box.



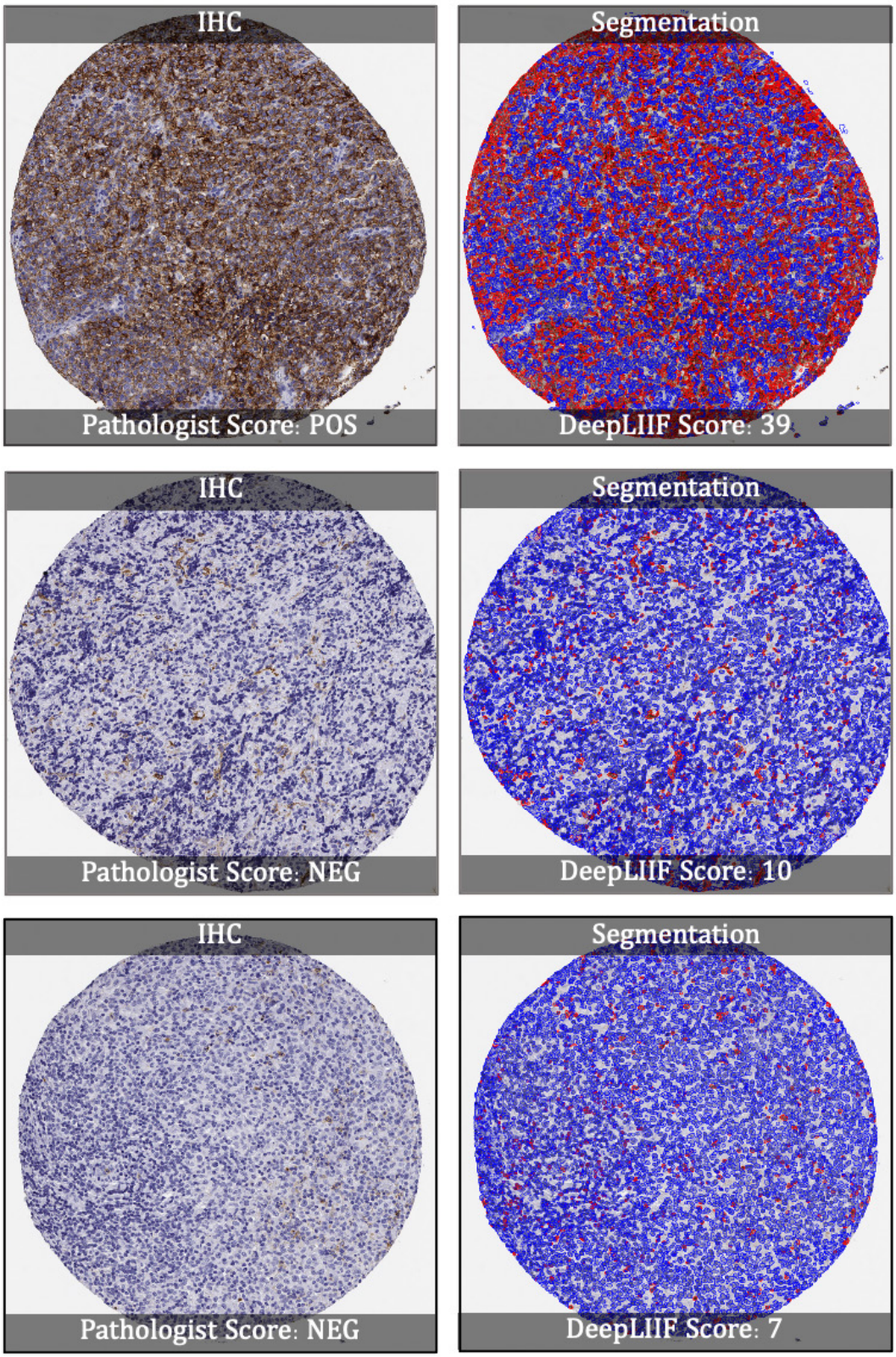
Extended Data Fig. 8. Some examples from LYON19 Challenge Dataset (27). The generated modalities and classified segmentation mask for each sample are shown in a separate box.



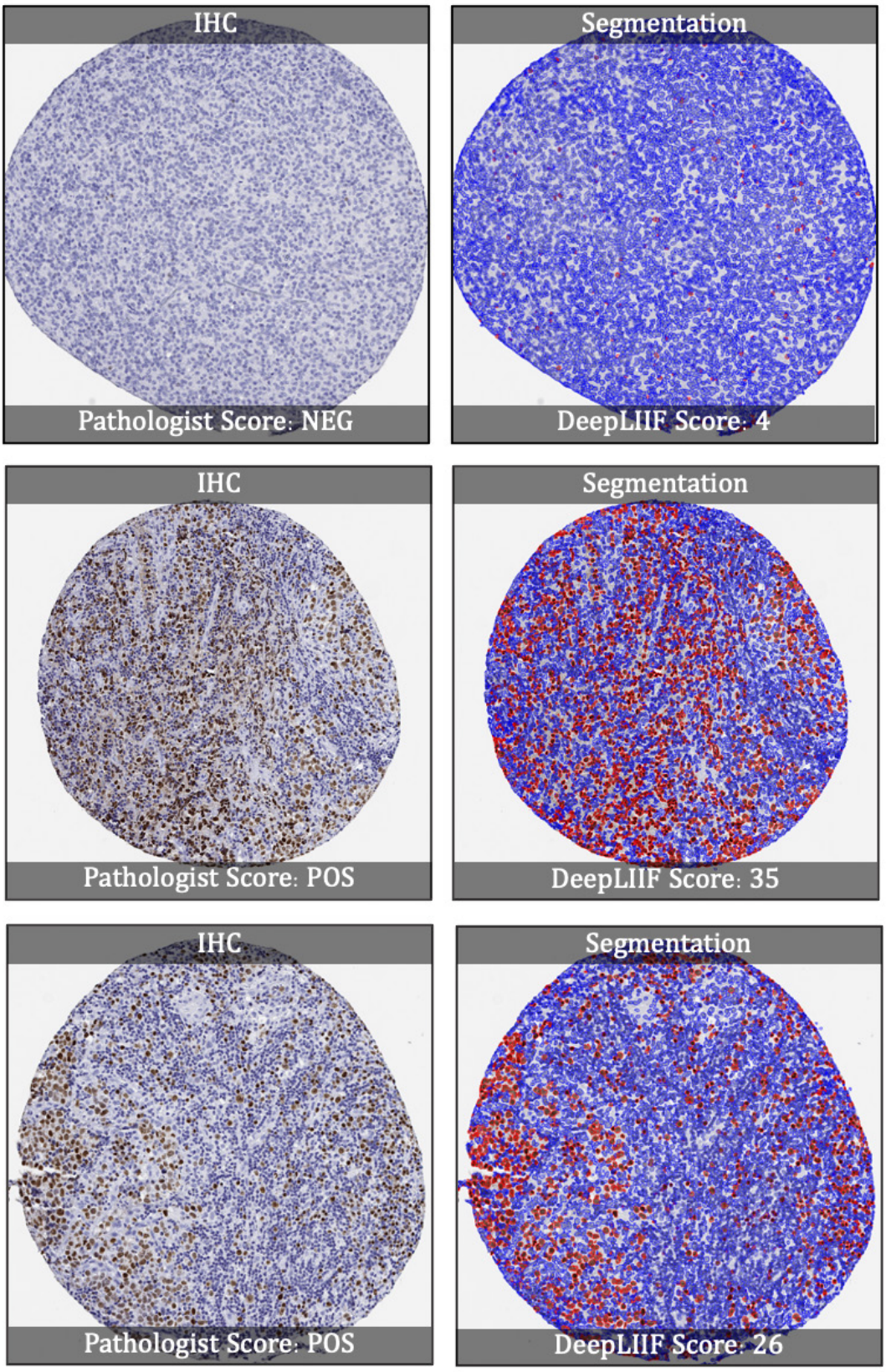
Extended Data Fig. 9. Some sample tissues stained with BCL2 marker from DLBCL-morph Dataset (34). In each row, the original IHC tissue image is shown on the left side, and the corresponding segmentation mask is shown on the right side.



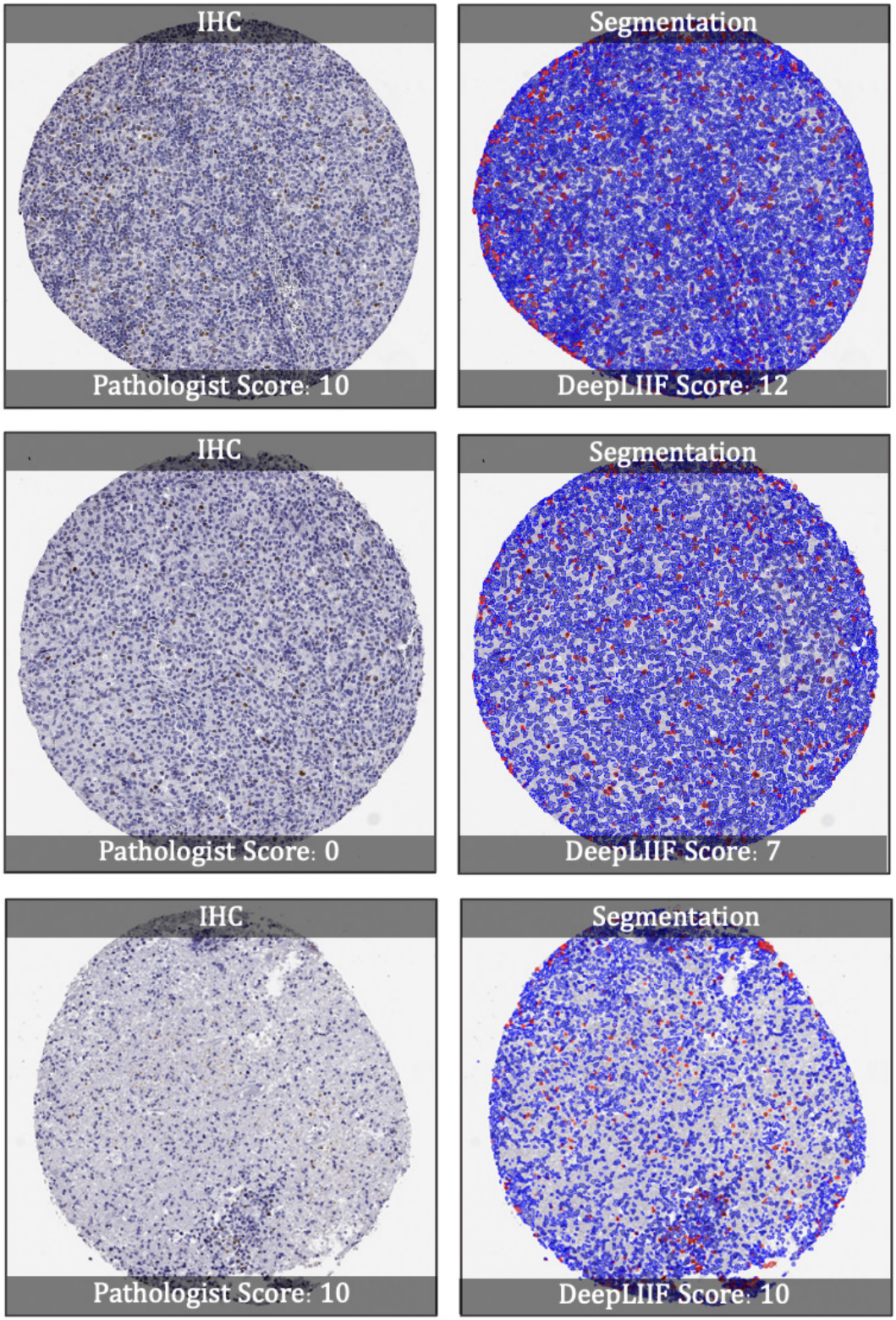
Extended Data Fig. 10. Some sample tissues stained with BCL6 marker from DLBCL-morph Dataset (34). In each row, the original IHC tissue image is shown on the left side, and the corresponding segmentation mask is shown on the right side.



Extended Data Fig. 11. Some sample tissues stained with CD10 marker from DLBCL-morph Dataset (34). In each row, the original IHC tissue image is shown on the left side, and the corresponding segmentation mask is shown on the right side.



Extended Data Fig. 12. Some sample tissues stained with MUM1 marker from DLBCL-morph Dataset (34). In each row, the original IHC tissue image is shown on the left side, and the corresponding segmentation mask is shown on the right side.



Extended Data Fig. 13. Some sample tissues stained with MYC marker from DLBCL-morph Dataset (34). In each row, the original IHC tissue image is shown on the left side, and the corresponding segmentation mask is shown on the right side.