An AI-Ready Multiplex Staining Dataset for Reproducible and Accurate Characterization of Tumor Immune Microenvironment

Parmida Ghahremani[1], Joseph Marino[1], Juan Hernandez-Prera[2], Janis V. de la Iglesia[2], Robbert J. C. Slebos[2], Christine H. Chung[2], and Saad Nadeem[1(✉)]

[1] Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
nadeems@mskcc.org
[2] Moffitt Cancer Center, Tampa, FL 33612, USA
christine.chung@moffitt.org

Abstract. We introduce a new AI-ready computational pathology dataset containing restained and co-registered digitized images from eight head-and-neck squamous cell carcinoma patients. Specifically, the same tumor sections were stained with the expensive multiplex immunofluorescence (mIF) assay first and then restained with cheaper multiplex immunohistochemistry (mIHC). This is a first public dataset that demonstrates the equivalence of these two staining methods which in turn allows several use cases; due to the equivalence, our cheaper mIHC staining protocol can offset the need for expensive mIF staining/scanning which requires highly-skilled lab technicians. As opposed to subjective and error-prone immune cell annotations from individual pathologists (disagreement $> 50\%$) to drive SOTA deep learning approaches, this dataset provides objective immune and tumor cell annotations via mIF/mIHC restaining for more reproducible and accurate characterization of tumor immune microenvironment (e.g. for immunotherapy). We demonstrate the effectiveness of this dataset in three use cases: (1) IHC quantification of CD3/CD8 tumor-infiltrating lymphocytes via style transfer, (2) virtual translation of cheap mIHC stains to more expensive mIF stains, and (3) virtual tumor/immune cellular phenotyping on standard hematoxylin images. The dataset is available at https://github.com/nadeemlab/DeepLIIF.

Keywords: multiplex immuofluorescence · multiplex immunohistochemistry · tumor microenvironment · virtual stain-to-stain translation

P. Ghahremani, J. Marino, C. H. Chung, and S. Nadeem—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_68.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 704–713, 2023.
https://doi.org/10.1007/978-3-031-43987-2_68

1 Introduction

Accurate spatial characterization of tumor immune microenvironment is critical for precise therapeutic stratification of cancer patients (e.g. via immunotherapy). Currently, this characterization is done manually by individual pathologists on standard hematoxylin-and-eosin (H&E) or singleplex immunohistochemistry (IHC) stained images. However, this results in high interobserver variability among pathologists, primarily due to the large (> 50%) disagreement among pathologists for immune cell phenotyping [10]. This is also a big cause of concern for publicly available H&E/IHC cell segmentation datasets with immune cell annotations from single pathologists. Multiplex staining resolves this issue by allowing different tumor and immune cell markers to be stained on the same tissue section, avoiding any phenotyping guesswork from pathologists.

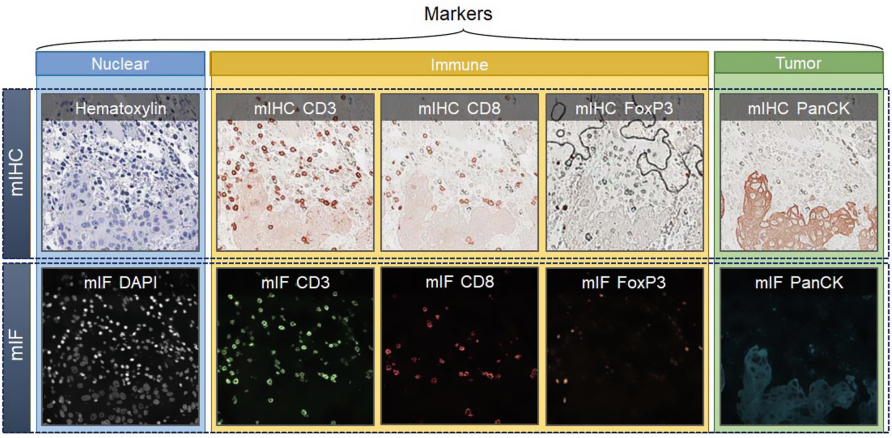

Fig. 1. Dataset overview. Restained and co-registered mIHC and mIF sample with nuclear (hematoxylin/DAPI), immune (CD3 - T-cell marker, CD8 - Cytotoxic T-cell, FoxP3 - regulatory T-cell), and tumor (PanCK) markers. CD3 = CD8 + FoxP3.

Multiplex staining can be performed using expensive multiplex immunofluorescence (mIF) or via cheaper multiplex immunohistochemistry (mIHC) assays. MIF staining (requiring expensive scanners and highly skilled lab technicians) allows multiple markers to be stained/expressed on the same tissue section (no co-registration needed) while also providing the utility to turn ON/OFF individual markers as needed. In contrast, current brightfield mIHC staining protocols relying on DAB (3,3'-Diaminobenzidine) alcohol-insoluble chromogen, even though easily implementable with current clinical staining protocols, suffer from occlusion of signal from sequential staining of additional markers. To this effect, we introduce a new brightfield mIHC staining protocol using alcohol-soluble aminoethyl carbazole (AEC) chromogen which allows repeated stripping, restaining, and scanning of the same tissue section with multiple markers. This

requires only affine registration to align the digitized restained images to obtain non-occluded signal intensity profiles for all the markers, similar to mIF staining/scanning.

In this paper, we introduce a new dataset that can be readily used out-of-the-box with any artificial intelligence (AI)/deep learning algorithms for spatial characterization of tumor immune microenvironment and several other use cases. To date, only two denovo stained datasets have been released publicly: BCI H&E and singleplex IHC HER2 dataset [7] and DeepLIIF singleplex IHC Ki67 and mIF dataset [2], both without any immune or tumor markers. In contrast, we release the first denovo mIF/mIHC stained dataset with tumor and immune markers for more accurate characterization of tumor immune microenvironment. We also demonstrate several interesting use cases: (1) IHC quantification of CD3/CD8 tumor-infiltrating lymphocytes (TILs) via style transfer, (2) virtual translation of cheap mIHC stains to more expensive mIF stains, and (3) virtual tumor/immune cellular phenotyping on standard hematoxylin images.

Table 1. Demographics and other relevant details of the eight anonymized head-and-neck squamous cell carcinoma patients, including ECOG performance score, Pack-Year, and surgical pathology stage (AJCC8).

ID	Age	Gender	Race	ECOG	Smoking	PY	pStage	Cancer Site	Cancer Subsite
Case1	49	Male	White	3	Current	21	1	Oral Cavity	Ventral Tongue
Case2	64	Male	White	3	Former	20	4	Larynx	Vocal Cord
Case3	60	Male	Black	2	Current	45	4	Larynx	False Vocal Cord
Case4	53	Male	White	1	Current	68	4	Larynx	Supraglottic
Case5	38	Male	White	0	Never	0	4	Oral Cavity	Lateral Tongue
Case6	76	Female	White	1	Former	30	2	Oral Cavity	Lateral Tongue
Case7	73	Male	White	1	Former	100	3	Larynx	Glottis
Case8	56	Male	White	0	Never	0	2	Oral Cavity	Tongue

2 Dataset

The complete staining protocols for this dataset are given in the accompanying **supplementary material**. Images were acquired at 20× magnification at Moffitt Cancer Center. The demographics and other relevant information for all eight head-and-neck squamous cell carcinoma patients is given in Table 1.

2.1 Region-of-Interest Selection and Image Registration

After scanning the full images at low resolution, nine regions of interest (ROIs) from each slide were chosen by an experienced pathologist on both mIF and mIHC images: three in the tumor core (TC), three at the tumor margin (TM), and three outside in the adjacent stroma (S) area. The size of the ROIs was standardized at 1356×1012 pixels with a resolution of 0.5 μm/pixel for a total surface area of 0.343 mm^2. Hematoxylin-stained ROIs were first used to align all

the mIHC marker images in the open source Fiji software using affine registration. After that, hematoxylin- and DAPI-stained ROIs were used as references to align mIHC and mIF ROIs again using Fiji and subdivided into 512×512 patches, resulting in total of 268 co-registered mIHC and mIF patches (∼33 co-registered mIF/mIHC images per patient).

2.2 Concordance Study

We compared mIF and mIHC assays for concordance in marker intensities. The results are shown in Fig. 2. This is the first direct comparison of mIF and mIHC

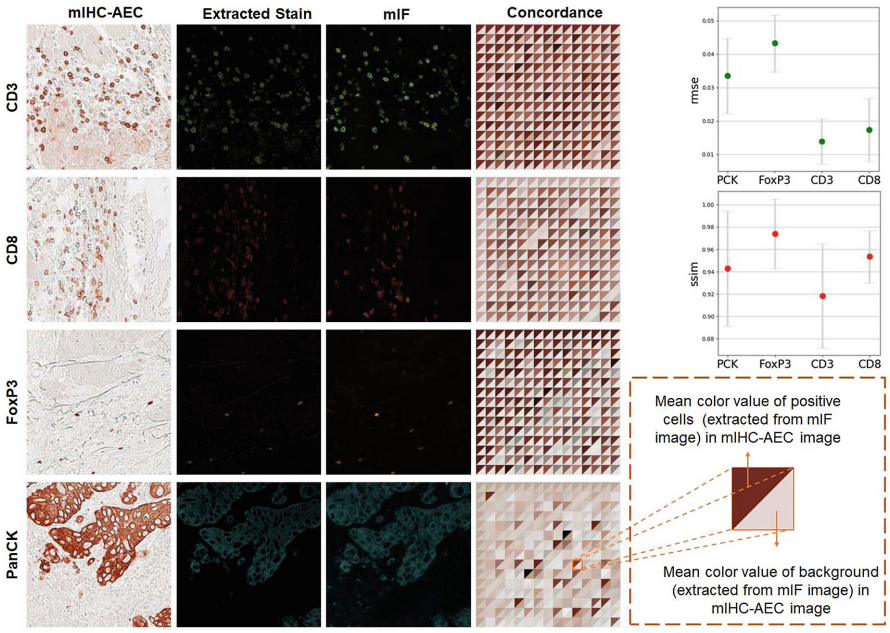

Fig. 2. Concordance Study. Second column shows stains extracted from first column mIHC-AEC images using Otsu thresholding. Third column shows the corresponding perfectly co-registered original mIF images. Using the mIF image, we separated foreground of the mIHC-AEC image from its background and calculated the mean value of the foreground pixels as well as the background pixels. The fourth column shows the results of the concordance study. Each square represents an image in the dataset and the top half of each square shows the mean color value of the positive cells, extracted from mIHC-AEC using its corresponding mIF image and the bottom half of it shows the mean color value of its background. *The high intensity of the top half of the squares represents positive cells and the low intensity of the bottom half represents non-positive cells (background), which is seen in almost all squares, demonstrating* ***high concordance among mIHC-AEC and mIF data***. The last column shows the RMSE and SSIM diagrams of all four stains calculated using the extracted stain from IHC-AEC images (second column) and the mIF images (third column). The low error rate of RMSE and high structural similarity seen in these diagrams show high concordance among mIHC-AEC and mIF images.

using identical slides. It provides a standardized dataset to demonstrate the equivalence of the two methods and a source that can be used to calibrate other methods.

3 Use Cases

In this section, we demonstrate some of the use cases enabled by this high-quality AI-ready dataset. We have used publicly available state-of-the-art tools such as Adaptive Attention Normalization (AdaAttN) [8] for style transfer in the IHC CD3/CD8 quantification use case and DeepLIIF virtual stain translation [2,3] in the remaining two use cases.

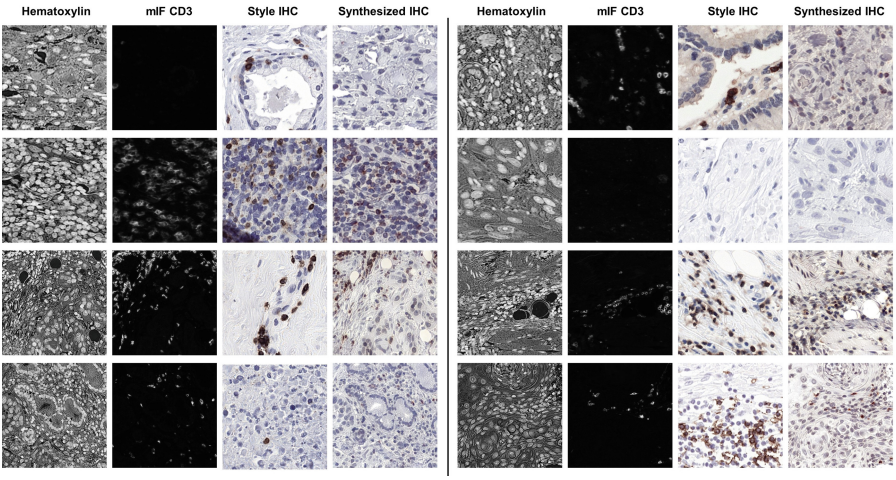

Fig. 3. Examples of synthesized IHC images and corresponding input images. Style IHC images were taken from the public LYON19 challenge dataset [14]. We used grayscale Hematoxylin images because they performed better with style transfer.

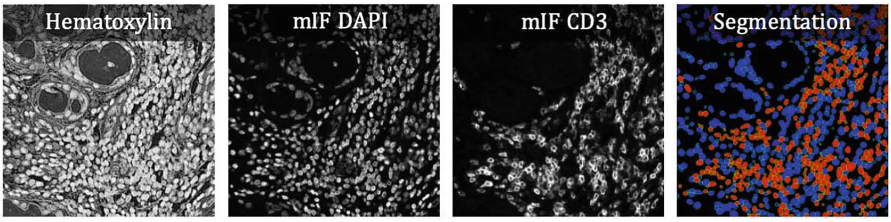

Fig. 4. Examples of Hematoxylin, mIF DAPI, mIF CD3 and classified segmentation mask for this marker. The DAPI images were segmented using Cellpose [13] and manually corrected by a trained technician and approved by a pathologist. The segmented masks were classified using the CD3 channel intensities.

Table 2. Quantitative metrics for NuClick and LYSTO testing sets. **F1** is the harmonic mean of recall and precision, **IOU** is intersection over union, and pixel accuracy (PixAcc) is $\frac{TP}{TP+FP+FN}$, where TP, FP, and FN represent the number of true positive, false positive, and false negative pixels, respectively.

Model	Dataset	NuClick			LYSTO
		F1↑	IOU↑	PixAcc↑	DiffCount↓
UNet [11]	NuClick	0.47 ± 0.30	0.36 ± 0.24	0.62 ± 0.37	10.06 ± 15.69
	Our Dataset	0.48 ± 0.29	0.36 ± 0.25	0.69 ± 0.37	2.91 ± 5.47
FPN [5]	NuClick	0.50 ± 0.31	0.39 ± 0.26	0.64 ± 0.38	2.82 ± 3.49
	Our Dataset	0.52 ± 0.31	0.40 ± 0.26	0.67 ± 0.36	1.90 ± 2.90
UNet++ [15]	NuClick	0.49 ± 0.30	0.37 ± 0.25	0.63 ± 0.37	2.75 ± 5.29
	Our Dataset	0.53 ± 0.30	0.41 ± 0.26	0.70 ± 0.36	2.19 ± 2.89

3.1 IHC CD3/CD8 Scoring Using mIF Style Transfer

We generate a stylized IHC image (Fig. 3) using three input images: (1) hematoxylin image (used for generating the underlying structure of cells in the stylized image), (2) its corresponding mIF CD3/CD8 marker image (used for staining positive cells as brown), and (3) sample IHC style image (used for transferring its style to the final image). The complete architecture diagram is given in the **supplementary material**. Specifically, the model consists of two sub-networks:

(a) Marker Generation: This sub-network is used for generating mIF marker data from the generated stylized image. We use a conditional generative adversarial network (cGAN) [4] for generating the marker images. The cGAN network consists of a generator, responsible for generating mIF marker images given an IHC image, and a discriminator, responsible for distinguishing the output of the generator from ground truth data. We first extract the brown (DAB channel) from the given style IHC image, using stain deconvolution. Then, we use pairs of the style images and their extracted brown DAB marker images to train this sub-network. This sub-network improves staining of the positive cells in the final stylized image by comparing the extracted DAB marker image from the stylized image and the input mIF marker image at each iteration.

(b) Style Transfer: This sub-network creates the stylized IHC image using an attention module, given (1) the input hematoxylin and the mIF marker images and (2) the style and its corresponding marker images. For synthetically generating stylized IHC images, we follow the approach outlined in AdaAttN [8]. We use a pre-trained VGG-19 network [12] as an encoder to extract multi-level feature maps and a decoder with a symmetric structure of VGG-19. We then use both shallow and deep level features by using AdaAttN modules on multiple layers of VGG. This sub-network is used to create a stylized image using the structure of the given hematoxylin image while transferring the overall color distribution of the style image to the final stylized image. The generated marker image from the first sub-network is used for a more accurate colorization of the

positive cells against the blue hematoxylin counterstain/background; not defining loss functions based on the markers generated by the first sub-network leads to discrepancy in the final brown DAB channel synthesis.

For the stylized IHC images with ground truth CD3/CD8 marker images, we also segmented corresponding DAPI images using our interactive deep learning ImPartial [9] tool https://github.com/nadeemlab/ImPartial and then classified the segmented masks using the corresponding CD3/CD8 channel intensities, as shown in Fig. 4. We extracted 268 tiles of size 512×512 from this final segmented and co-registered dataset. For the purpose of training and testing all the models, we extract four images of size 256×256 from each tile due to the size of the external IHC images, resulting in a total of 1072 images. We randomly extracted tiles from the LYON19 challenge dataset [14] to use as style IHC images. Using these images, we created a dataset of synthetically generated IHC images from the hematoxylin and its marker image as shown in Fig. 3.

We evaluated the effectiveness of our synthetically generated dataset (stylized IHC images and corresponding segmented/classified masks) using our generated dataset with the NuClick training dataset (containing manually segmented CD3/CD8 cells) [6]. We randomly selected 840 and 230 patches of size 256×256 from the created dataset for training and validation, respectively. NuClick training and validation sets [6] comprise 671 and 200 patches, respectively, of size 256×256 extracted from LYON19 dataset [14]. LYON19 IHC CD3/CD8 images are taken from breast, colon, and prostate cancer patients. We split their training set into training and validation sets, containing 553 and 118 images, respectively, and use their validation set for testing our trained models. We trained three models including UNet [11], FPN [5], UNet++ [15] with the backbone of resnet50 for

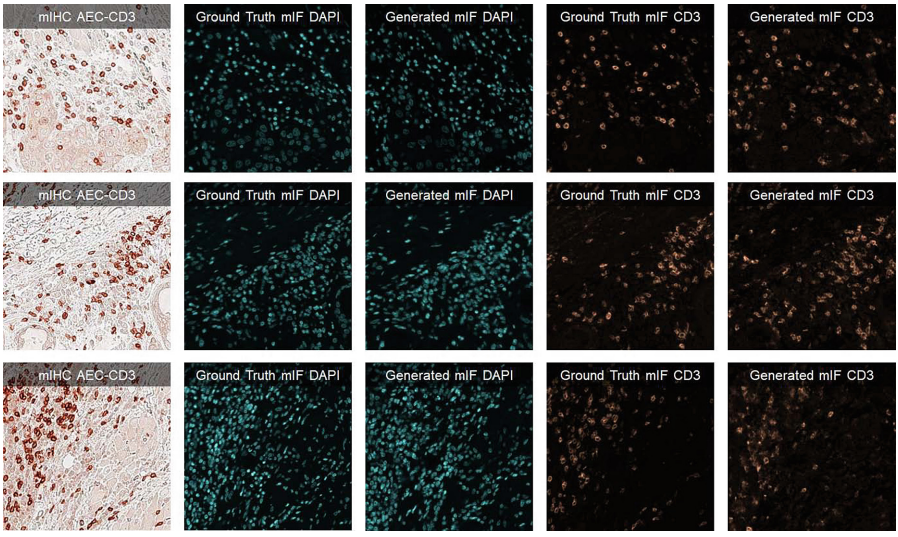

Fig. 5. Examples of ground-truth and generated mIF data from mIHC-AEC images.

200 epochs and early stopping on validation score with patience of 30 epochs, using binary cross entropy loss and Adam optimizer with learning rate of 0.0001. As shown in Table 2, models trained with our synthetic training set outperform those trained solely with NuClick data in all metrics.

We also tested the trained models on 1,500 randomly selected images from the training set of the Lymphocyte Assessment Hackathon (LYSTO) [1], containing image patches of size 299×299 obtained at a magnification of $40\times$ from breast, prostate, and colon cancer whole slide images stained with CD3 and CD8 markers. Only the total number of lymphocytes in each image patch are reported in this dataset. To evaluate the performance of trained models on this dataset, we counted the total number of marked lymphocytes in a predicted mask and calculated the difference between the reported number of lymphocytes in each image with the total number of lymphocytes in the predicted mask by the model. In Table 2, the average difference value (**DiffCount**) of lymphocyte number for the whole dataset is reported for each model. As seen, the trained models on our dataset outperform the models trained solely on NuClick data.

3.2 Virtual Translation of Cheap mIHC to Expensive mIF Stains

Unlike clinical DAB staining, as shown in style IHC images in Fig. 3, where brown marker channel has a blue hematoxylin nuclear counterstain to stain for all the cells, our mIHC AEC-stained marker images (Fig. 5) do not stain for all the cells including nuclei. In this use case, we show that mIHC marker images can be translated to higher quality mIF DAPI and marker images which

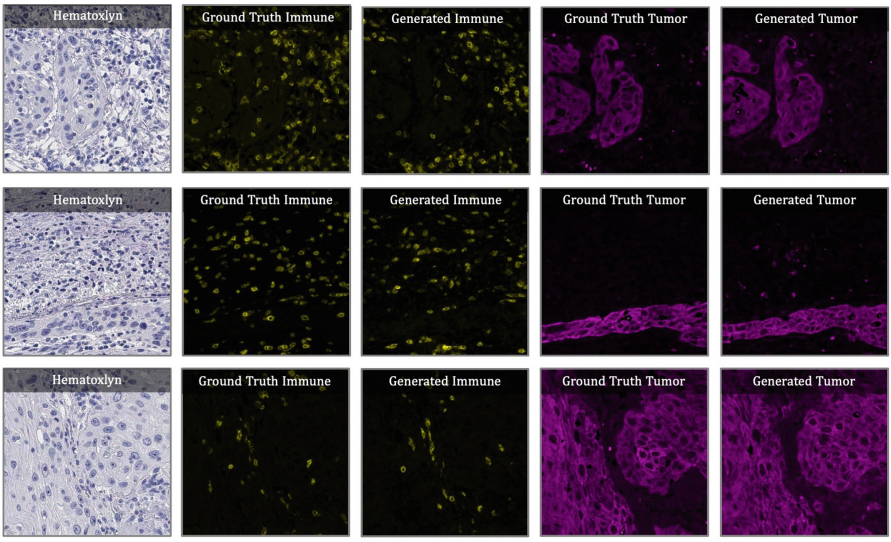

Fig. 6. Examples of ground-truth and generated mIF immune (CD3) and tumor (PanCK) markers from standard hematoxylin images.

stain effectively for all the cells. We used the publicly available DeepLIIF virtual translation module [2,3] for this task. We trained DeepLIIF on mIHC CD3 AEC-stained images to infer mIF DAPI and CD3 marker. Some examples of testing the trained model on CD3 images are shown in Fig. 5. We calculated the Mean Squared Error (MSE) and Structural Similarity Index (SSIM) to evaluate the quality of the inferred modalities by the trained model. The MSE and SSIM for mIF DAPI was 0.0070 and 0.9991 and for mIF CD3 was 0.0021 and 0.9997, indicating high accuracy of mIF inference.

3.3 Virtual Cellular Phenotyping on Standard Hematoxylin Images

There are several public H&E/IHC cell segmentation datasets with manual immune cell annotations from single pathologists. These are highly problematic given the large ($> 50\%$) disagreement among pathologists on immune cell phenotyping [10]. In this last use case, we infer immune and tumor markers from the standard hematoxylin images using again the public DeepLIIF virtual translation module [2,3]. We train the translation task of DeepLIIF model using the hematoxylin, immune (CD3) and tumor (PanCK) markers. Sample images/results taken from the testing dataset are shown in Fig. 6.

4 Conclusions and Future Work

We have released the first AI-ready restained and co-registered mIF and mIHC dataset for head-and-neck squamous cell carcinoma patients. This dataset can be used for virtual phenotyping given standard clinical hematoxylin images, virtual clinical IHC DAB generation with ground truth segmentations (to train high-quality segmentation models across multiple cancer types) created from cleaner mIF images, as well as for generating standardized clean mIF images from neighboring H&E and IHC sections for registration and 3D reconstruction of tissue specimens. In the future, we will release similar datasets for additional cancer types as well as release for this dataset corresponding whole-cell segmentations via ImPartial https://github.com/nadeemlab/ImPartial.

Data use Declaration and Acknowledgment: This study is not Human Subjects Research because it was a secondary analysis of results from biological specimens that were not collected for the purpose of the current study and for which the samples were fully anonymized. This work was supported by MSK Cancer Center Support Grant/Core Grant (P30 CA008748) and by James and Esther King Biomedical Research Grant (7JK02) and Moffitt Merit Society Award to C. H. Chung. It is also supported in part by the Moffitt's Total Cancer Care Initiative, Collaborative Data Services, Biostatistics and Bioinformatics, and Tissue Core Facilities at the H. Lee Moffitt Cancer Center and Research Institute, an NCI-designated Comprehensive Cancer Center (P30-CA076292).

References

1. Ciompi, F., Jiao, Y., Laak, J.: Lymphocyte assessment hackathon (LYSTO) (2019). https://zenodo.org/record/3513571
2. Ghahremani, P., et al.: Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. Nat. Mach. Intell. **4**, 401–412 (2022)
3. Ghahremani, P., Marino, J., Dodds, R., Nadeem, S.: Deepliif: an online platform for quantification of clinical pathology slides. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21399–21405 (2022)
4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
5. Kirillov, A., He, K., Girshick, R., Dollár, P.: A unified architecture for instance and semantic segmentation (2017). http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf
6. Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N.: NuClick: a deep learning framework for interactive segmentation of microscopic images. Med. Image Anal. **65**, 101771 (2020)
7. Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M.: BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix (Accepted CVPR Workshop). arXiv preprint arXiv:2204.11425 (2022)
8. Liu, S., et al.: AdaAttN: revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6649–6658 (2021)
9. Martinez, N., Sapiro, G., Tannenbaum, A., Hollmann, T.J., Nadeem, S.: Impartial: partial annotations for cell instance segmentation. bioRxiv, pp. 2021–01 (2021)
10. Reisenbichler, E.S., et al.: Prospective multi-institutional evaluation of pathologist assessment of pd-l1 assays for patient selection in triple negative breast cancer. Mod. Pathol. **33**(9), 1746–1752 (2020)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods **18**(1), 100–106 (2021)
14. Swiderska-Chadaj, Z., et al.: Learning to detect lymphocytes in immunohistochemistry with deep learning. Med. Image Anal. **58**, 101547 (2019)
15. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1