Virtual Immunohistochemistry Staining for Histological Images Assisted by Weakly-supervised Learning

Supplementary Material

HCC is a type of liver cancer, and it accounts for about 80% of all liver cancer cases. In the clinical diagnosis, doctors will first cut out a piece of tissue from the liver. Subsequently, the tissue is made into sections stained with H&E, and a preliminary judgment of the presence of liver cancer is made based on the H&E staining reaction. However, the diagnosis of HCC cannot be fully confirmed based solely on the results of H&E staining. Therefore, the doctor will cut the adjacent tissue layers of the H&E stained slides and process the tissue into IHC stained slides through immunohistochemical reactions. Finally, further diagnosis of HCC will be conducted based on the results of the IHC staining reaction.

It is frustrating that the generation of IHC-stained sections through chemical staining takes a long time, making it unsuitable for rapid processing. Therefore, using virtual staining technology to generate IHC staining images is an urgently needed problem to be solved. GPC3 is currently the most promising protein to be considered as a specific tumor marker for HCC, and is widely recognized as an early marker for liver cancer and a target for immunotherapy, attracting extensive attention from doctors.

In order to advance the diagnosis of HCC, we have collected the first H&E-GPC3 virtual staining dataset named HCI. All data comes from Peking University Shenzhen Hospital, and the project is approved by the ethical commission of Peking University Shenzhen Hospital (2023089). Next we will provide a detailed introduction to the collection and preparation process of this dataset.

5.1. Collection

The data scanning equipment is SQS-600P, a pathology section scanner with a scanning speed of 15 seconds per slice. In the end, we selected 30 H&E slides and 30 GPC3 slides for training, 10 H&E slides and their corresponding GPC3 slides for validation, as well as 10 H&E slides and their corresponding GPC3 slides for testing. The 100 slides were eventually divided into non-overlapping patches of size 256×256. Additionally, we use the pos/neg expression of GPC3 protein in the pathological report as the WSI level label for H&E slides.

5.2. Preparation

Images Registration. Although H&E-GPC3 slides have been collected in the test set and valid set, the pixel-level differences are significant in H&E-GPC3 image pairs because these data are all from adjacent layers of tissue. In order to better validate the performance of different virtual staining algorithms, we used a registration algorithm to align them.

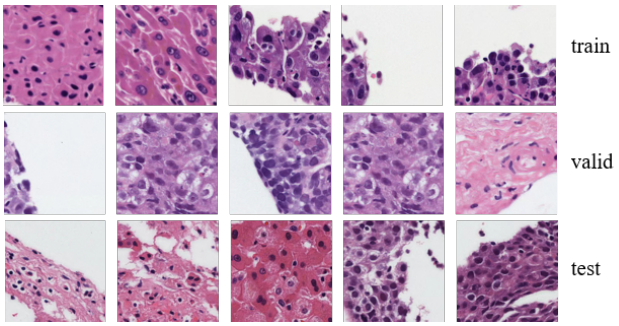

Figure 6. The dataset do not undergo stain normalization. As shown in the figure, there are significant color differences among different patches.

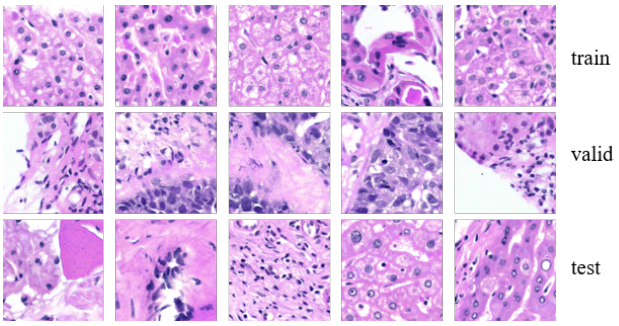

Figure 7. The data set consist of images that have been normalized for staining. It can be seen that the color differences among the images are reduced.

Stain Normalization. Due to different preservation times, as well as variations in the amount of dye used during staining, there may be some differences among different slices, as shown in Fig. 6.

From the Fig. 6, it can be seen that there is a significant color difference between different patches in the dataset. These color variations may cause the staining network to pay too much attention to color information and overlook more important information such as cellular structure.

To deal with this problem, we performed stain normalization on all patches in the HCI dataset, as shown in Fig. 7. It can be seen that after stain normalization, the color differences between patches can be significantly reduced.