Nuclear Segmentation and Classification Model with Imbalanced Classes for CoNiC Challenge

Jijun Cheng†, Xipeng Pan†, Feihu Hou, Bingchao Zhao, Jiatai Lin, Zhenbing Liu, Zaiyi Liu, Chu Han

Abstract—Nuclear segmentation and classification is an essential step for computational pathology. TIA lab from Warwick University organized a nuclear segmentation and classification challenge (CoNiC) for H&E stained histopathology images in colorectal cancer based on the Lizard dataset [1]. In this challenge, computer algorithms should be able to segment and recognize six types of nuclei, including Epithelial, Lymphocyte, Plasma, Eosinophil, Neutrophil, Connective tissue. This challenge introduces two highly correlated tasks, nuclei segmentation and classification task and prediction of cellular composition task. There are a few obstacles we have to address in this challenge, 1) imbalanced annotations with few training samples on minority classes, 2) color variation of the images from multiple centers or scanners, 3) limited training samples, 4) similar morphological appearance among classes. To deal with these challenges, we proposed a systematic pipeline for nuclear segmentation and classification. First, we built a GAN-based model to automatically generate pseudo images for data augmentation. Then we trained a self-supervised stain normalization model to solve the color variation problem. Next we constructed a baseline model HoVer-Net with cost-sensitive loss to encourage the model pay more attention on the minority classes. According to the results of the leaderboard, our proposed pipeline achieves 0.40665 mPQ+ (Rank 33rd) and 0.62199 r2 (Rank 4th) in the preliminary test phase.

Index Terms—Nuclear segmentation and classification, Cost-sensitive, Data Augmentation, Imbalanced Annotations

I. INTRODUCTION

Graham (et al.) from the TIA lab (Warwick University) organized a nuclei segmentation and classification challenge (CoNiC) on Grand Challenge Platform based on the recently proposed Lizard dataset [1], which is the largest nuclei segmentation and classification dataset for colorectal cancer with 495,179 nuclear instances with classification labels. There are two tasks in this challenge, the segmentation and classification task and the cellular composition task. Different from the other existing datasets like Pannuke [2], Lizard only focuses on colorectal cancer and it further divides inflammatory nuclei into several sub-categories, lymphocyte, plasma, neutrophil or eosinophil. There are still a lot of obstacles have to be solved.

1) The classification labels are extremely imbalanced. This challenge comprised of 54,850 lymphocytes, 9,821 plasma cells, 3,666 neutrophils and 2,775 eosinophils. The number of lymphocytes is almost 20 times larger than the number of eosinophils, as shown in Fig. 1. Lack of training samples make neural network models hard to train a well feature representation for the minority classes. 2) Since the images are from multiple centers in TCGA, color variation cannot be ignored because of different staining protocols or scanners. 3) Even it is the largest dataset so far for only one specific cancer, we can say the samples are still not enough for training a robust model because of the heterogeneous of malignant tumor. 4) The morphological appearances among different types are similar, such as lymphocytes and plasma cells, tumor epithelial nuclei and normal epithelial nuclei. How to solve the aforementioned challenges might be the key to build a stable and general model for this challenge.

In this paper, we proposed a systematic pipeline for CoNiC challenge. It includes several individual steps, including data augmentation, stain normalization and segmentation and classification. First, We applied our previously proposed GAN-based model [3] to retrain a data augmentation model specifically on Lizard dataset and generated over 1000 pseudo images. Next, we utilized our proposed self-supervised stain normalization model, which also retrained on Lizard dataset, to eliminate the color variation. Finally, we applied a strong baseline model HoVer-Net as the backbone with a cost-sensitive loss to further balance the classes during network training. We can see all the steps have been proven its effectiveness...
according to the quantitative results on the leaderboard in the preliminary test phase. Finally, our proposed pipeline achieves 0.40665 mPQ+ (Rank 33rd) and 0.62199 r2 (Rank 4th) in the preliminary test phase.

II. METHODOLOGY

Here we demonstrate our proposed nuclear segmentation and classification pipeline with following three part: data augmentation, stain normalization and the baseline model.

A. Data Augmentation

Preparing rich training samples are always the most important portion in the deep learning era, even more important than constructing a complex model. Limited training data makes it difficult to train deep learning model for nuclear segmentation and classification. So we introduced a previous GAN-based work of our research group [3] for data augmentation. This model use the masks from Lizard dataset as the input image and generate fake histopathology images. With a well trained generator, we randomly construct new pseudo labels and generated pseudo pathological images. Fig. 2 shows the process of generating pathological images.

In order to avoid inputting exactly the same masks to the generator while keeping the real distribution of the nuclei, we introduce some very slight vibrations for each real mask. First of all, we extract every nuclear instance in the semantic segmentation masks in the Lizard dataset. For each nuclear instance, we introduced random $x$ and $y$ axis translations within the range of $(-3, 3)$ pixels and a random rotation within the angle of $-3^\circ$ and $+3^\circ$ while keeping the class and the shape unchanged. Such slight changes aim to conforms to the real distribution of cells in the real histopathological images. We generated 1000 pseudo image and mask pairs for data augmentation. Fig. 3 shows the cases of generating pseudo pathological images using pseudo labels.

B. Stain normalization

Color variation is also one of the greatest challenges in computational pathology. The staining differences caused by multiple centers and across tumors not only affect the training of the model but also hurt its generalization ability. Since this challenge does not allow to use other public pathology datasets. We applied a self-supervised stain normalization approach RestainNet [4] we proposed recently. We used the CoNSeP dataset, which is a part of Lizard dataset, to train our stain normalization model. Since RestainNet is a self-supervised model, paired training images are not necessary, which shows great flexibility in preparing training data. By applying RestainNet, all the images were mapped into the same staining style with CoNSeP dataset.

C. Network Architecture

To the best of our knowledge, HoVer-Net [5] is the strongest baseline model for nuclear segmentation and classification. Thanks to the horizontal and vertical distance maps design, HoVer-Net is able to separate the overlapping nuclei. So in this challenge, HoVer-Net was employed as our baseline method, which was proposed to segment and classify nuclei in histology images simultaneously. The network consisted of three branches: Nuclear Pixel (NP), HoVer, Nuclear Classification (NC). NP was designed for nuclear semantic segmentation, and the HoVer learned the horizontal and vertical map of the nuclei. Combined with the results of NP, the markers were available for watershed algorithm that was used for post-processing. NC predicted the category of nuclear pixels, and merged the segmentation results of the above two branches to obtain the final results. Besides the original losses introduced in HoVer-Net, we also proposed a cost-sensitive loss in the NC branch.

1) Cost-sensitive loss: Imbalanced data is a regular occurrence in target identification and classification tasks. The imbalance of different cell types is particularly pronounced in the CoNiC challenge. For example, epithelials, lymphocytes and connective cells contribute for a substantial portion of the ratio, whereas eosinophils and neutrophils account for a negligible portion. The imbalance of cell types will result in a decrease in segmentation and classification performance. We proposed a cost-sensitive loss function for multi-class nuclear segmentation and classification to address this problem. By incorporating the cost-sensitive matrix into the loss function, the impact of sample imbalance could be alleviated, and the classification accuracy of small categories can be improved. For multi-class tasks, it is assumed that the number of categories is $N$. The cost-sensitive matrix $M$ with the size of
N × N can be defined as follows:

\[
    M = \begin{bmatrix}
        m_{(0,0)} & m_{(0,1)} & \cdots & m_{(0,N-1)} \\
        m_{(1,0)} & m_{(1,1)} & \cdots & m_{(1,N-1)} \\
        \vdots & \vdots & \ddots & \vdots \\
        m_{(N-1,0)} & m_{(N-1,1)} & \cdots & m_{(N-1,N-1)}
    \end{bmatrix}
\]

where \( m_{(j,k)} \) represents the cost when class \( k \) is incorrectly classified as class \( j \). The diagonal element of matrix \( M \) is 0. The values of non-diagonal elements on the cost-sensitive matrix are set according to the ratio of the number of samples.

III. RESULTS

In the preliminary test phase, the results of segmentation and classification are shown in Table I and the cellular composition results are shown in Table II. Note that, we did not specifically design a regression model for the task 2, the results of the task 2 are directly generated from the results of the task 1. According to the results of the leaderboard, the mPQ+ and r2 of our proposed pipeline achieves 0.40665(Rank 33rd) and 0.62199(Rank 4th), respectively. We can observe that both stain normalization and data augmentation can greatly benefit the nuclear segmentation and classification task for all the classes.

ACKNOWLEDGMENT

This research was supported in part by National Science Fund for Distinguished Young Scholars, China (Grant No. 81925023), National Natural Science Foundation of China (Grant Nos. 62002082, 62102103), China Postdoctoral Science Foundation (Grant No. 2021M690753), 2020 National Innovation and Entrepreneurship Training Program for College Students (Grant No. 202010595023).

REFERENCES

[1] S. Graham, M. Jahanifar, A. Azam, M. Nimir, Y.-W. Tsang, K. Dodd, E. Hero, H. Sahota, A. Tank, K. Benes et al., “Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 684–693.