I. Methods

A. Mask R-CNN

To address the constraint of limited training data, we have opted for the adoption of a lightweight Mask R-CNN [2] architecture, leveraging a pretrained model for fine-tuning on the competition dataset. The schematic representation of the Mask R-CNN architecture is illustrated in Figure 1. The input image undergoes initial processing through a residual network [3][4] to compute deep features. Subsequently, these deep features are combined using a Feature Pyramid Network (FPN) across various scales to obtain more informative deep features, enabling the network to detect objects of different sizes.

Following this, the Region Proposal Network (RPN) utilizes the feature map generated by FPN to identify regions that may contain foreground objects. Initially, the RPN generates nine different anchors with three different aspect ratios and three different scales for each grid cell in the feature map. Using softmax, the probability of each anchor belonging to a foreground object is calculated, and high-probability anchors are selected to produce Region of Interest (ROI) proposals, which are potential foreground object regions. Subsequently, ROI Align reuses the feature map generated by FPN to extract features corresponding to the ROI proposals. For non-integer coordinates in the proposals, ROI Align employs interpolation based on surrounding integer coordinates, ensuring more accurate segmentation map generation compared to rounding the coordinates.

Finally, each ROI proposal undergoes a three-branch decision head for instance segmentation. Two branches initially detect the object's bounding box and determine its class, while the third branch generates segmentation maps for all possible classes, from which the map corresponding to the specific class is selected. Separating segmentation and classification effectively mitigates competition among different classes at object boundaries, ensuring more comprehensive segmentation of object regions. Consequently, Mask R-CNN enables precise detection and segmentation of every object in the image.

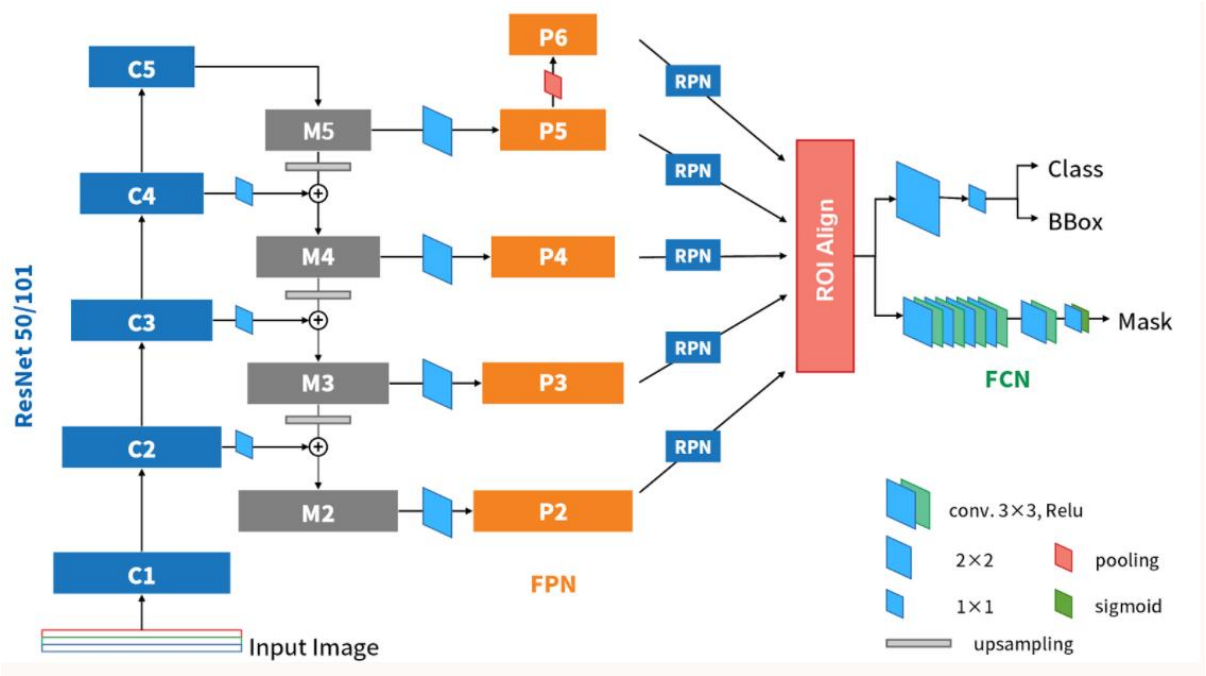

圖 1、Mask-RCNN Structure (Source: https://ivan-eng-murmur.medium.com/99370c98de28)

B. Model Ensembling

Different models may capture distinct features, with one model potentially recognizing aspects that another may overlook. Therefore, combining detection results from multiple diverse models is likely to enhance overall detection accuracy. We attempted to train two distinct models and then employed non-maximum suppression (NMS) to integrate the detection outcomes of both models. In instances where there is a high degree of overlap in detected regions, NMS retains the prediction with the highest confidence from each model, eliminating redundant detections and improving the overall robustness of the detection results.

C. Image Normalization

During the model training process, to mitigate the impact of varying data value scales and enhance training efficiency, we perform normalization on the input images. This involves subtracting the mean of the dataset (approximately [127, 127, 127]) and dividing by the standard deviation of the dataset (approximately [15, 15, 15]). This normalization procedure ensures that the

numerical values of the input images are uniformly adjusted to follow a distribution with a mean of 0 and a standard deviation of 1.

Additionally, we explored an alternative approach by directly employing the mean ([103.53, 116.28, 123.675]) and standard deviation ([57.375, 57.12, 58.395]) from the ImageNet dataset for normalization purposes. This experimentation aims to evaluate the impact of using different normalization parameters on the training process and model performance.

D. Augmentation

During the model training process, to prevent overfitting to a fixed set of training images and to expose the model to a broader range of image variations for learning more diverse features, we employ random horizontal or vertical flips, as well as random scaling, to introduce geometric augmentation. This involves randomly applying transformations such as flips and scaling to the training images. The augmented data is then used to train the model.

In addition to geometric augmentation, we also experiment with randomly adjusting the brightness, contrast, and saturation of the training images, thereby altering the color to create various color augmentations. This approach aims to further enhance the variability of the training data, allowing the model to better generalize across different color representations and improving its overall robustness.

E. Test Time Augmentation

Similar to model ensembling, when images transform, a model may detect different aspects. Therefore, by generating multiple augmentations of test images and combining the detection results from all augmentations, detection accuracy can be enhanced. We flip and scale test images to create all possible geometric augmentations that may have occurred during training. Subsequently, we utilize non-maximum suppression (NMS) to integrate the detection results of the model across all augmentations. This approach ensures a comprehensive evaluation of the model's performance under various geometric transformations, contributing to improved detection accuracy.

II. Reference

1. Sartorius, “Sartorius - Cell instance segmentation,” *Kaggle,* 2021.
2. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Intnl. Conf. Computer Vision,* pp. 2961-2969, 2017.
3. K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” *IEEE Conf. Computer Vision and Pattern Recognition*,” 2016.
4. S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He, “Aggregated Residual Transformations for Deep Neural Networks,” *IEEE Conf. Computer Vision and Pattern Recognition*,” 2017.