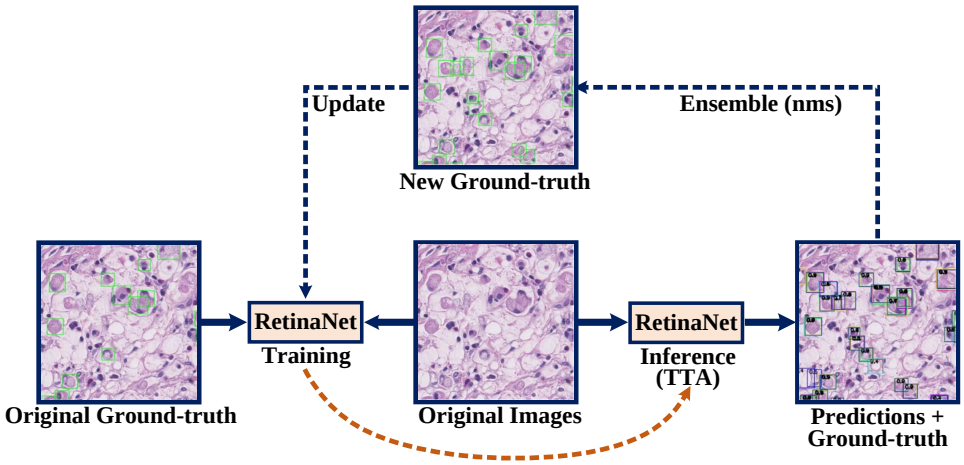

Fig. 6: The pipeline of cascaded self-training detection network developed by the zju_realdoctor team.

Supplementary Material

7. Solutions of Top Teams

We provide detailed solutions of each top-5 ranked teams in two challenges. Each solution include two parts, i.e., methodology overview with framework figures and implementation details.

7.1. Task-1: Signet Ring Cell Detection

7.1.1. zju_realdoctor (Qingyu Song)

Overview: To achieve the detection of signet ring cells, Song develops a cascaded self-training detection network with a modified RetinaNet (Lin et al., 2018). Fig. 6 presents their developed pipeline. The pipeline first trains a RetinaNet on original images based on the given ground-truth annotations. Considering the incomplete annotations of signet ring cells in the given data, it then employs the trained RetinaNet to infer the training images by using test time augmentation (TTA) to get as many missed labeled signet ring cells as possible. After the inference of training images, a modified Non-Maximum Suppression (NMS) strategy is introduced to generate new training labels, which will be used in the next round of training. The pipeline can be iteratively trained until there are no improvements in the valid recall.

Implementation Details: Different from the original RetinaNet, Song modifies the RetinaNet to make it learns from the normal images. When it comes to a positive image, RetinaNet will set anchors which have IoU greater than 0.5 with the label of positive and less than 0.2 with the label of negative. Accordingly, all anchors in normal regions are set as negative, which can learn from negative images by RetinaNet. In addition to the detection part, the framework employs a classification part that focuses on decreasing the number of false-positive in normal regions. An image is considered positive in the final classification if the number of predicted bounding boxes is larger than a pre-defined number.

In the network training, the zju_realdoctor team randomly crops 512×512 regions from original large-sized training images with up to 2000×2000. For the unbalanced data of positive and negative problems, the pipeline employs a resampling strategy to make the number of crop regions in positive and negative images the same in each epoch. In the detection part, for the inference of original images, test time augmentation (TTA) is employed, including fliplr, flipud, rotation with 90^o, 180^o, 270^o. To fuse these predictions and ground-truth bounding boxes, a modified non-maximum suppression is employed, which has the original function of NMS and an extra function to eliminate the bounding boxes with a lower score if another one includes 80% area of a bounding

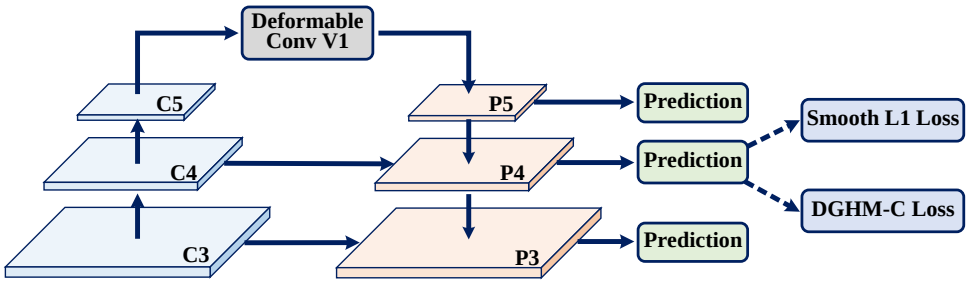

Fig. 7: The pipeline for signet ring cell detection developed by the SJTU_MedicalCV team, namely Dconv-RetinaNet

box. Besides, a voting strategy is used to get positive and negative results, i.e., if more than or equal to 2 models, classify the image into positive after 4-fold cross-validation. Finally, the zju_realdoctor team achieved the 1^{st} place in the Task-1 challenge, with the Recall of 0.8774, FPNormal of 100.00 and the FROC of 0.8774.

7.1.2. SJTU_MedicalCV (Jiancheng Yang et al.)

Overview: As presented in Fig. 7, Yang et al. develop Dconv-RetinaNet, which employs the RetinaNet as the network architecture and adopt ImageNet pre-trained ResNet as the backbone with Feature Pyramid Network (FPN) structure. Particularly, a deformable convolutional layer is added for better feature extraction on the top of feedforward ResNet. They employ Smooth L1 loss for regression and Decoupled GHM-C loss for classification, where "Decoupled" indicates that the anchors are divided from different labels in different images and computed the loss respectively. In addition to the Dconv-RetinaNet, Faster RCNN and Cascade RCNN are also used for the ensemble.

Implementation Details: The SJTU_MedicalCV team adopts ResNet18 as the backbone, demonstrating better performance than ResNet50 and EfficientNet. As the overcrowded regions of signet ring cells may be unlabeled, they decouple the classification loss functions into three parts (i.e., anchors in negative images, negative anchors in positive images, and positive anchors in positive images), which embed the GHM loss and handle the outliers respectively. For the regression loss, only anchors assigned to ground-truth is computed, where Smooth L1 loss function is used for simplicity. In the ensemble stage, three post-processing strategies are introduced for a whole image predicted by RetinaNet: 1) if the number of bounding boxes is less than 30, it is considered to be negative, where all such images are dropped; 2) if the number of bounding boxes is more than 50, it is considered to be positive, where the bounding boxes are concatenated by the predictions of 3 deep models and then applied with NMS strategy; 3) if the number of bounding boxes is between 30 and 50, the confidence is set in the sigmoid without changing their ranking. For the data augmentation, the original 2000×2000 images are decomposed into 800×800 patches with the stride of 300 pixels, followed by random horizontal flipping and random crop to the scale of 600 pixels. To balance the positive and negative samples during training, positive-negative sample ratios are manually set as $1:3$ when constructing a mini-batch.

7.1.3. mirl_task1 (Sreehari S et al.)

Overview: Fig. 8 presents the pipeline developed by the mirl_task1 team. Given the original histopathological images, a patch-based approach is first employed for detection, which is necessary for extending the solution on whole-slide images. Subsequently, Mask R-CNN is employed for detection, which involves a multi-resolution feature extraction block, region proposal block, followed by classifier head. ResNet-X is the backbone for feature extraction and collects output from various intermediate layers as multi-resolution features. Afterward, the multi-resolution features are used by the region proposal network, which regresses for bounding boxes and confidence value estimation. The proposed regions are gradient updates via the classifier and regressor loss. The model

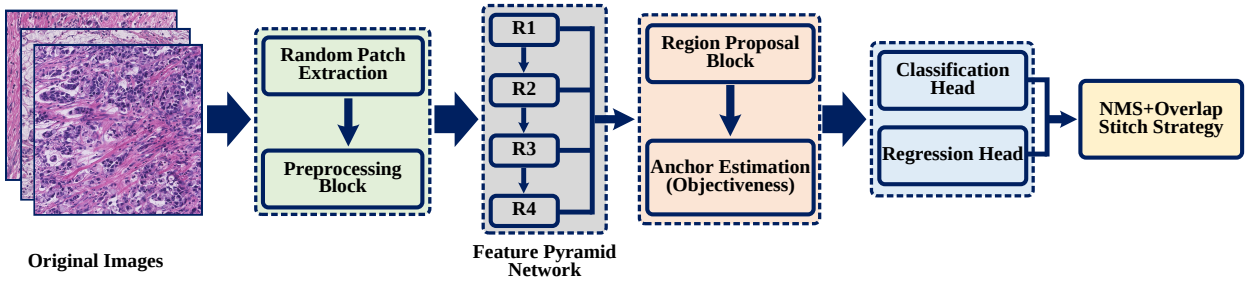

Fig. 8: The pipeline for the signet ring cell detection developed by the mirl_task1 team, where the above case R1, R2, R3 and R4 denotes features from multiple dimensions.

is also trained on negative images using a hard-mining strategy.

Implementation Details: ResNet-X 101 (He et al., 2016) and Feature Pyramid Network (FPN) (Lin et al., 2017a) are employed in the Mask R-CNN network, which serves as a feature extractor. From the region proposal predictions, top anchors with higher foreground scores are picked, whereas Non-Maximal Suppression (NMS) is used to avoid too many overlapping boxes. The model is trained on 512×512 sized patches. Data augmentation strategies include rotation, translation, re-scaling, horizontal and vertical flipping, and color jitter. Each patch is normalized to have zero mean and unit variance, followed by min-max normalization to bound the pixel values in the range of $0 - 1$. In the training stage, the model is first trained with given positive data, with cross-entropy and squared error loss functions with Adam optimizer. To further train the model on the negative sample, the labels are extracted due to false positives predicted by the model trained on positive examples only. Finally, the model is trained with all the positive samples and hard mined negative samples. In addition, multiple inferences and post-processing strategies are implemented. The tissue mask of the given pathological images is estimated and extracted, which reduces the number of operations. For the patch extraction, patches are sequentially extracted with overlapping strides to address edge effects. Then overlapping bounding box predictions are removed by region thresholding.

7.1.4. szucv517 (Weizeng Lu et al.)

Overview: Fig. 9 presents the pipeline developed by the szucv517 team. The pipeline consists of two important parts, i.e., a detector and a refinement filter. The detector contains two detection modules: 1) Dual Shot Face Detector (DSFD) and PyramidBox. The outputs of DSFD and pyramid box are fused by voting for more robust detection results. Considering the large number of negative cells that are similar to the signet ring cells, a refinement filter is cascaded at the end of the detector, i.e., a binary classifier that utilizes SE-ResNet18 as the backbone network. The voting results of the bounding boxes are used to crop cells from images, which is taken as the input of the refinement filter. After predicting the bounding boxes of each patch, the final result of the original image is obtained according to the voting function.

Implementation Details: As the signet ring cell dataset has a large number of dense targets with overlapping areas, DSFD and Pyramidbox are employed as the baseline. For the false positives in the output of the detector, it is necessary to use background information to filter out the false positives. The refinement filter is a SE-ResNet18 classification network cascaded at the end of the detector to filter the false positives. The product of classification confidence and detection confidence is set as the final confidence of the bounding box. In the implementation, only positive images are used for the detector training. Each image is cropped into a number of 640×640 patches using a sliding window with a step size of 128, which generates $11,2976$ annotated patches for training. The VGG-16 network pre-trained on ImageNet is adopted as the backbone of DSFD and Pyramidbox. Subsequently, the trained detector is fed with 378 negative images and gets bounding boxes with confidence above 0.15. As the signet ring cells in

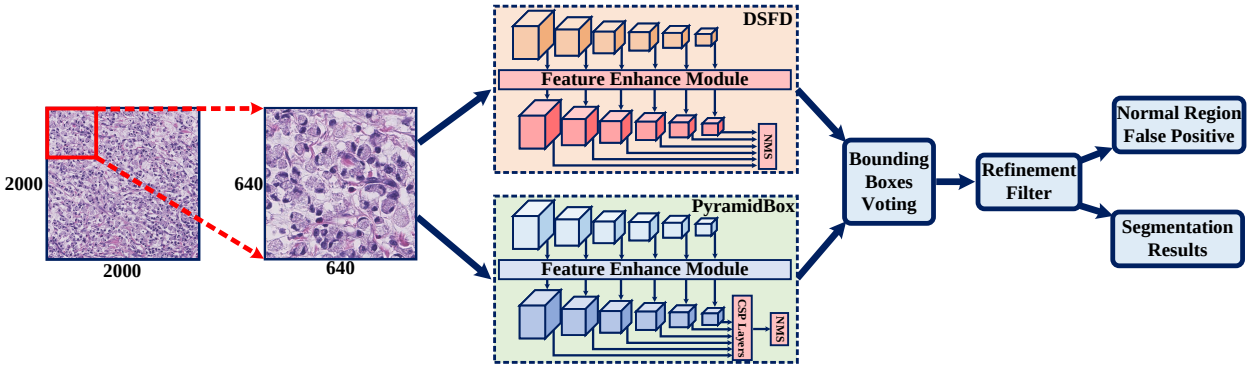

Fig. 9: The pipeline for the signet ring cell detection developed by the szucv517 team.

positive images are not fully annotated, the detector is also fed with 77 positive images to get all bounding boxes with confidence above 0.15.

7.1.5. HFUTB906 (Jun Shi et al.)

Overview: YOLO v3-SPP (Redmon and Farhadi, 2018) is employed as the core algorithm in HFUTN906 team. Particularly, the SPP module is applied in YOLO v3 (Redmon and Farhadi, 2018) for the extraction of multi-scale deep features with different receptive fields. Then the deep features are fused by concatenating in the channel dimension of feature maps. For the unlabeled positive samples in the dataset, since the lack of a semi-supervised approach in the current pipeline, the team does not use this part of the data, which is not labeled.

Implementation Details: For the large-sized histopathological images, the sliding window strategy is used to produce training samples. Considering the limitations of hardware devices and the efficiency of training, the window size is fixed. The SPP module is consists of 1 convolutional layer and three parallel max-pooling layers with the kernel sizes of 5×5, 9×9, and 13×13. The module is integrated with YOLO v3 between the 5th and the 6th convolutional layers in front of the first detection header. All the data produced is used for training.

7.2. Task-2: Colonoscopy Tissue Segmentation and Classification

7.2.1. kuangkuang (Ke Mei et al.)

Overview: As presented in Fig. 10, Mei et al. develop a three-stage classification and segmentation pipeline that classifies and segments top-down. For the whole-slide images, Stage-1 is the classification for WSI level, judging whether a WSI is benign or malignant, with the output of the classification result of WSI. Subsequently, Stage-2 is the classification in the patch level, i.e., classifying each image patch in the malignant WSI. Finally, Stage-3 is the segmentation of positive patches, then stitching patches into a complete WSI mask.

Implementation Details: The sliding window (stride = 512, size = 1536×1536) is first employed to crop WSI images. All deep models are trained with patch-level images. Online data augmentation strategy is employed, including random folds, random brightness contrast, and grid distortion. In Stage-1, positive patches and negative patches are sampled equally from both positive WSIs and negative WSIs for training. The classifier is the DenseNet-161 with ImageNet pre-trained parameters (Huang et al., 2017). When inferring a WSI, all the WSI patches are first classified and get a set of classification results. If the patch classified as positive accounts for more than 20% of all patches, the WSI is considered positive. The average of all positive or negative patches' scores is used for this WSI. In Stage-2, considering the fact that the Stage-1 classifier performer poorly (81.32% accuracy) on patches from

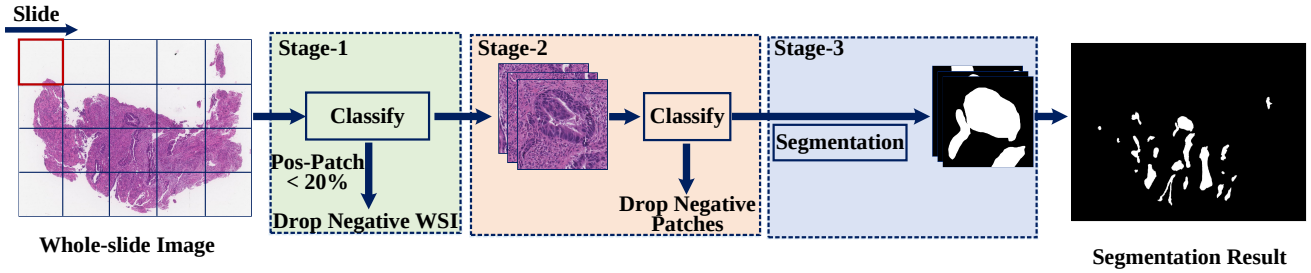

Fig. 10: The pipeline for the colonoscopy tissue segmentation and classification developed by the *kuangkuang* team.

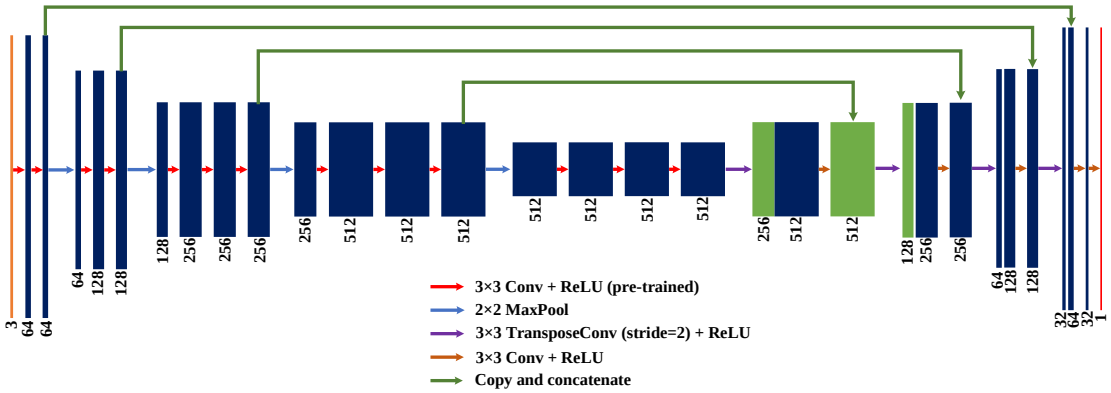

Fig. 11: The network architecture used by the team zju_realdoctor, which employs the pre-trained convolutional layer of VGG-16 as the U-Net encoder.

the same positive WSI, all patches from the positive WSI are selected as training data. Three models (i.e., ResNext10 (Xie et al., 2017), InceptionV3 (Szegedy et al., 2016), and DenseNet161 (Huang et al., 2017)) are used for integration. In Stage-3, for the positive patches found in Stage-2, a segmentation network is trained based on the U-Net with ResNet50 as the encoder. In addition, several other modules are added to improve the performance, including IBM block, pyramid pooling module (PPM), spatial-channel squeeze & excitation (SCSE) block, etc.

7.2.2. zju_realdoctor (Xuechen Liu)

Overview: The zju_realdoctor team presents a sliding window-based framework with fine-tuned U-Net. As shown in Fig. 11, a pre-trained VGGNet is employed as U-Net's encoder. In the training stage, square image patches are randomly cropped with the same size to fine-tune the U-Net. During the inference, image patches are cropped at equal intervals using a sliding window. A pathological slice is split by a sliding window, where patches are fed into the network to generate segmentation masks. The malignant probability of the slice can be obtained based on the segmentation results.

Implementation Details: In the preprocessing stage, four folds cross-validation data are the first split based on the stratified sampling to balance the benign and malignant samples. To increase the receptive field of the model, the original pathological images are down-sampled to 1/4 size, which means 512 patch size has the 2048 receptive field on the original image. For the network architecture, the last layer of downsampling in the VGGNet is removed. Then the first four layers of convolution are used as the encoder of the network. During the training, the benign and malignant samples are simultaneously fed to the model to obtain a well-trained model that can suppress the benign sample's output. Random rotation, random vertical and horizontal flipping are used for the data augmentation. Additionally, a window size of 2048 is employed to generate patches from the test image. A threshold of 0.2 is selected to generate the final binary mask from the probability map.

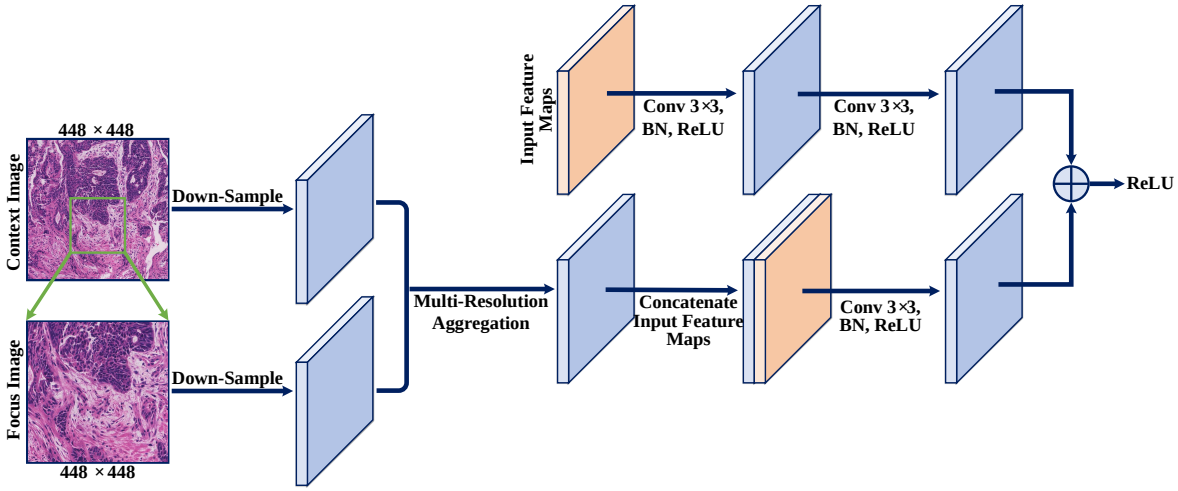

Fig. 12: The pipeline for the colonoscopy tissue segmentation and classification developed by the *TIA_Lab* team mainly includes the context-aware MIL unit that re-introduces the multi-resolution image into the network after the max-pooling operation.

7.2.3. TIA_Lab (Simon Graham et al.)

Overview: Simon et al. develop a fully convolutional neural network that utilizes a multi-resolution input to provide additional context. As demonstrated in Fig. 12, the approach extends MILD-Net (Graham et al., 2019) with a new minimal information loss residual unit that re-introduces the multi-resolution input at various points within the network to counter the loss of information caused by max-pooling and to provide additional context. For the network's output, segmentation is performed at each image scale, where the contextual patch prediction acts as an auxiliary output to encourage the network to learn contextual information.

Implementation Details: For the context-aware input, two image patches are extracted centered at a point on the original WSI of dimensions 448×448 and 896×896. Then, the larger patch is down-sampled to the size 448×448 by bicubic interpolation so that the size of both image patches is equal. For the network architecture, instead of using a single-resolution image at the input to the network and with the MIL unit, a multi-resolution input is employed that consists of two image patches, i.e., focus image and context image. At the network's output, the malignant regions are segmented in both the focus image and the context image. However, only the focus image segmentation results are considered during processing. To perform image classification, the team first smooths the raw probability map by applying a Gaussian kernel, and then the maximum probability of the entire image is extracted. In the experiment, Adam optimization with a batch size of 8 is used with an initial learning rate of $10(-4)$. Multiple data augmentation strategies are employed, including flip, rotation, shear, scale, and elastic transformations. Besides, blur, Gaussian noise, and color augmentations are also applied to improve model generalization to new data.

7.2.4. SJTU_MedicalCV (Canqian Yang et al.)

Overview: The SJTU_MedicalCV team develops specialized models for the two sub-tasks, i.e., image-wise classification and segmentation, respectively, and trains the two tasks individually on the same data. During inference, the two trained models are cascaded to serve as the overall inference pipeline. Particularly, given a large spatial resolution whole-slide image (WSI), the sliding window method is first employed to decompose the WSI into several patches with overlaps. The classifier determines the probabilities of the input patches belonging to a positive instance. These patches with high classification scores will be fed into the downstream network for further fine-grained segmentation.

Implementation Details: The ImageNet pre-trained EfficientNet (Tan and Le, 2019) is employed as the backbone network for the classification model. WSIs are decomposed into patches by a proposed method similar to multiple instance learning (Maron and

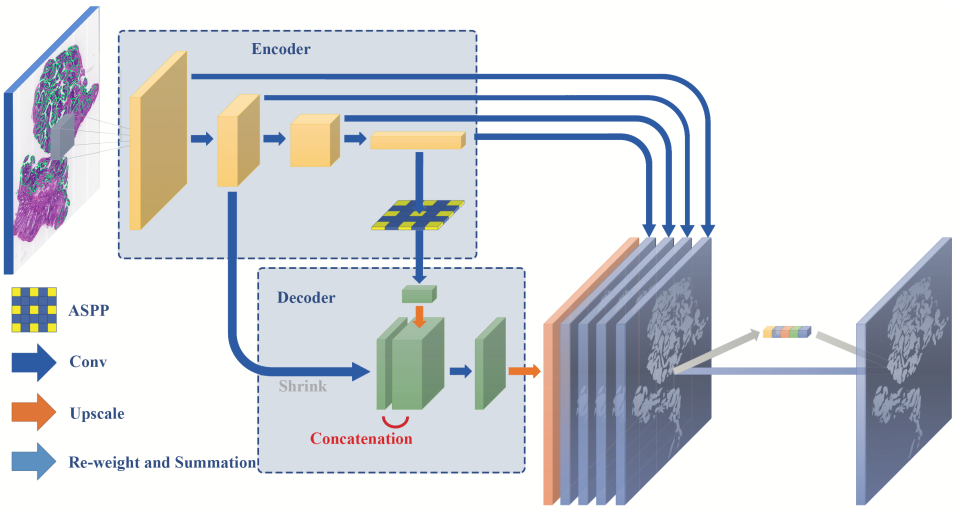

Fig. 13: Overview of segmentation model proposed by the SJTU_MedicalCV team.

Lozano-Pérez, 1998; Carbonneau et al., 2018), which can achieve better results in this task instead of using the exact label which can be obtained from the given segmentation mask. Besides, the team also develops a segmentation model, as depicted in Fig. 13. The model follows the encoder-decoder architecture of the DeepLabV3+ (Chen et al., 2018). Particularly, the team introduces deep supervision (DSV) (Lee et al., 2015) to enhance features extracted from the backbone network at all levels. Also, the Spatial and Channel Squeeze & Excitation (scSE) module (Roy et al., 2018) is employed to re-weight those features and generate the final segmentation. For the preprocessing, the original $3,000$-pixel images are decomposed into $1,024$-pixel patches with stride of $0.9\times$ patch size. The balance sample strategy is applied to handle highly unbalanced classes in the dataset. Common data augmentation strategies are used, including random flipping horizontally or vertically, random scaling in the range of $[0.5, 1.0]$, random rotation by 90^{o} and random crop of 512×512 image patches.

7.2.5. ustc_czw (Zhuowei Chen et al.)

Overview: The ustc_czw team presents their solution using the baseline model of PSP-Net (Zhao et al., 2017) with the backbone of ResNet50 (He et al., 2016). This selection is based on the experimental evaluation of the Dice value. Besides the baseline model, the team ignores doubtful annotations when training models, removing noisy samples. Subsequently, two models are trained, i.e., a model trained with original annotations, another model trained with the noiseless label by ignoring doubtful samples. Then the probability maps predicted by two models are averaged as the final probability map. The most significant value on the probability map is viewed as the results for classification, where the binary mask for segmentation is generated by the probability map with a 0.5 threshold.

Implementation Details: Firstly, the team randomly splits $520/130$ tissue slices as a training/validation dataset. In the training set, roughly $40,000$ patches can be obtain with 512×512 pixel size, including $10,000$ positive patches and $30,000$ negative patches. In the training phase, the input patch size is fixed as 512×512. The patches are trained on 3 different magnification rates ($10\times$, $20\times$, $40\times$) respectively, where $20\times$ is considered as the best magnification rate according to the context information and fine details. All negative examples are added into the training set. Meanwhile, the ratio of positive to negative examples is controlled to 1 in every training batch, which can enrich the diversity in the training set.