A Weakly Supervised Segmentation Network Embedding Cross-Scale Attention Guidance and Noise-Sensitive Constraint for Detecting Tertiary Lymphoid Structures of Pancreatic Tumors

Bingxue Wang, Liwen Zou, Jun Chen, Yingying Cao, Zhenghua Cai, Yudong Qiu, Liang Mao, Zhongqiu Wang, Jingya Chen, Luying Gui, and Xiaoping Yang

***Abstract*—The presence of tertiary lymphoid structures (TLSs) on pancreatic pathological images is an important prognostic indicator of pancreatic tumors. Therefore, TLSs detection on pancreatic pathological images plays a crucial role in diagnosis and treatment for patients with pancreatic tumors. However, fully supervised detection algorithms based on deep learning usually require a large number of manual annotations, which is time-consuming and labor-intensive. In this paper, we aim to detect the TLSs in a manner of few-shot learning by proposing a weakly supervised segmentation network. We firstly obtain the lymphocyte density maps by combining a pretrained model for nuclei segmentation and a domain adversarial network for lymphocyte nuclei recognition. Then, we establish a cross-scale attention guidance mechanism by jointly learning the coarse-scale features from the original histopathology images and fine-scale features from our designed lymphocyte density attention. A noise-sensitive constraint is introduced by an embedding signed distance function loss in the training procedure to reduce tiny prediction errors. Experimental results on two collected datasets demonstrate that our proposed method significantly outperforms the state-of-the-art segmentation-based algorithms in terms of TLSs detection accuracy. Additionally, we apply our method to study the congruent relationship between the density of TLSs and peripancreatic vascular invasion and obtain some clinically statistical results.**

***Index Terms*—Cross-scale attention, noise-sensitive constraint, pancreatic tumor, tertiary lymphoid structures, weakly supervised segmentation.**

I. Introduction

HISTOPATHOLOGY images are often the gold standard for disease detection, diagnosis, and prognostic analysis. However, analysis for such images is a very challenging task because of their large sizes and numerous elements. In recent years, the development of deep neural networks has led to many breakthroughs in automatic histopathology image classification and segmentation [1], [2]. These methods depend highly on extensive training and accurate pixel-level labels, which are particularly labor-intensive and time-consuming due to the huge sizes of pathological images. Actually, only a trained pathologist can distinguish precisely among the elements. Therefore, acquiring accurate annotations on pathological images is more complicated than annotating other medical images. The development of weakly supervised and unsupervised learning methods has become an inevitable trend in pathological image processing [3], [4].

In pathology research, the immune micro-environment is a topic of great concern. The main research target of the immune micro-environment is lymphocytes. Their particular aggregation clusters, known as tertiary lymphoid structures (TLSs), have received significant attention in recent years. Actually TLSs are discrete structured tissues of infiltrating immune cells, or in other words, more organized aggregation structures of lymphocytes. In some studies, TLSs are considered as any lymphoid aggregate similar to secondary lymphoid organs in non-lymphoid

Manuscript received 2 June 2023; revised 20 October 2023, 28 November 2023, and 1 December 2023; accepted 3 December 2023. Date of publication 8 December 2023; date of current version 6 February 2024. This work was supported in part by the China's Ministry of Science and Technology under Grant SQ2020YFA0713800, in part by the National Natural Science Foundation of China under Grants 12090023, 11971229, and 12001273, and in part by the Fundamental Research Funds for the Central Universities under Grant 30922010904. *(Corresponding authors: Xiaoping Yang; Luying Gui.)*

Bingxue Wang and Luying Gui are with the School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: bxwang@njust.edu.cn; ly.gui@njust.edu.cn).

Liwen Zou and Xiaoping Yang are with the Department of Mathematics, Nanjing University, Nanjing 210093, China (e-mail: dz20210008@smail.nju.edu.cn; xpyang@nju.edu.cn).

Jun Chen is with the Department of Pathology, Nanjing Drum Tower Hospital, Nanjing 210008, China (e-mail: ichenjun@qq.com).

Yingying Cao, Zhongqiu Wang, and Jingya Chen are with the Department of Radiology, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing 210000, China (e-mail: 983630231@qq.com; zhongqiuwang@njucm.edu.cn; chenjy@njucm.edu.cn).

Zhenghua Cai is with Medical School, Nanjing University, Nanjing 210007, China (e-mail: wxeyczh@163.com).

Yudong Qiu and Liang Mao are with the Department of General Surgery, Nanjing Drum Tower Hospital, Nanjing 210008, China (e-mail: yudongqiu510@nju.edu.cn; maoliang@njglyy.com).

Digital Object Identifier 10.1109/JBHI.2023.3340686

structures [5]. Several researches [6], [7] concentrated on the presence, composition, and location of TLSs in health and disease tissues and organs. The high density of TLSs found in many cancers is associated with prolonged survival of patients, such as colorectal cancers [8], [9] and breast cancers [10].

Previous studies have generally considered pancreatic tumors as cold tumors and lack of immune response [11], [12]. In contrast, some studies in recent years have found that the immune micro-environment of pancreatic tumors is highly heterogeneous, which largely contributes to its lack of mechanical clarity to date. One of the critical elements of its heterogeneity is a large number of immune cell subsets [13]. This means accurately segmenting lymphocytes and classifying clusters are crucial for studying pancreatic tumors. In the pathologist's opinion, lymphocytes in a TLS have a particular specific aggregation pattern, such as having a distinctive high density. Although a TLS does not have an envelope but is still clearly delimited from the surrounding tissues. However, it is challenging to identify all TLSs on a pathological image because of the considerable size variants of TLSs and their irregular distributions on the histopathology images. In addition, backgrounds with different colors and textures may interfere with the recognition of TLSs by humans. Also, an essential feature of the pancreas lacking immune response determines that there are fewer lymphocytes and TLSs in the pancreas, and it is more difficult to obtain a large number of training samples.

In order to accurately detect TLSs on a small-scale dataset containing whole slide images (WSIs) of hematoxylin and eosin (H&E) stained histopathology, we propose a weakly supervised segmentation network embedding cross-scale attention guidance and noise-sensitive constraint in this work. The proposed method is mainly divided into three parts: (1) all lymphocyte nuclei are segmented and identified from the unlabeled pancreatic pathology images by combining a pretrained nuclei segmentation model and adversarial learning. The corresponding lymphocyte density maps are constructed; (2) based on the aggregation characteristics of TLSs, we convert the lymphocyte density map in WSI to the attention guidance and establish a cross-scale attention network to learn the TLS features from different scales; (3) considering that the location of a TLS is critical information, we introduce a noise-sensitive constraint by a signed distance function to train a segmentation network with weak bounding box annotation. It is designed to make the network explicitly learn the TLSs' location distribution. To verify the effectiveness of our method in few-shot learning, the proposed method is trained on a small-scale dataset and evaluate on another large-scale dataset. The experimental results show that our method can accurately detect TLSs, and outperforms the state-of-the-art (SOTA) segmentation-based algorithms in terms of detection accuracy. As a clinical application, we use the density of TLSs to predict peripancreatic vascular invasion based on our proposed method.

The major contributions of this work are listed as follows.

- To the best of our knowledge, it is the first work to realize the few-shot detection task of TLSs on pancreatic histopathology images by embedding cross-scale attention guidance and noise-sensitive constraint into a weakly supervised segmentation network, which can be applied to the TLS analysis for other tumors.
- We construct the density map of lymphocytes which reflects the distribution characteristics of TLSs on pathological images and propose a cross-scale attention mechanism by jointly learning the TLS features from the coarse-scale pathological images and the fine-scale lymphocyte density maps.
- A noise-sensitive constraint with a signed distance function loss is introduced for training the TLS segmentation network with weak bounding box annotations, which helps to explicitly learn the TLS distribution and avoid enormous performance drops caused by tiny predicted errors.
- Experimental results show that our method significantly outperforms the state-of-the-art (SOTA) segmentation-based algorithms in terms of the TLS detection accuracy on pancreatic pathological images. And we validate that the TLS density is significantly related to the peripancreatic vascular invasion by our proposed method based on the clinical data acquired from two independent institutions.

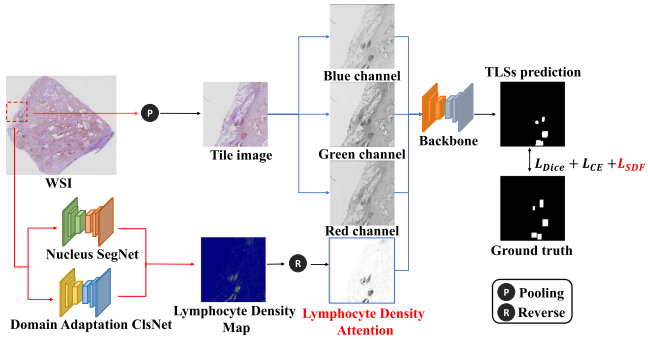

Fig. 1. Pipeline of our proposed segmentation network embedding cross-scale attention guidance and noise-sensitive constraint for weakly supervised TLSs segmentation of pancreatic cancer.

II. Method

We propose a weakly supervised segmentation network for the few-shot detection of TLSs on pancreatic pathological images by embedding cross-scale attention guidance and noise-sensitive constraint to improve the TLS segmentation performance. The pipeline of our proposed method is shown in Fig. 1. There are three core modules in our proposed model: (1) Lymphocyte nuclei segmentation and classification by combining nuclei segmentation baseline model and domain adversarial learning; (2) cross-scale attention guidance mechanism by jointly learning the TLS features from the coarse-scale H&E image and the fine-scale lymphocyte density map; (3) a noise-sensitive constraint by embedding a sign distance function loss for training the segmentation network with weak bounding box annotations.

When the proposed model is fed with an H&E WSI from a patient with pancreatic tumor, there are two processing flows for extracting cross-scale features of the TLS targets. The original H&E image is fed into a pretrained baseline model for nuclei segmentation and a domain adversarial network for lymphocyte nuclei classification. Then, we construct the lymphocyte density

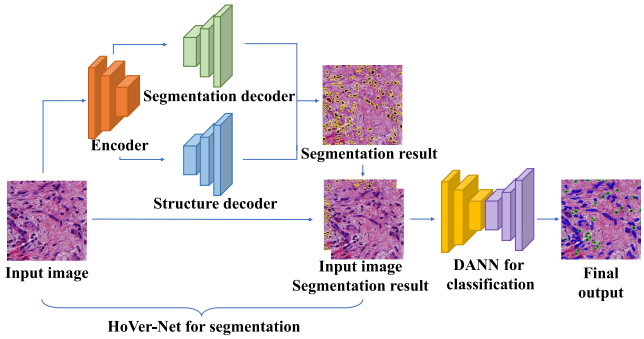

Fig. 2. Structure of the designed combination of Hover-Net for nuclei segmentation and domain adversarial neural network (DANN) for lymphocyte nuclei classification.

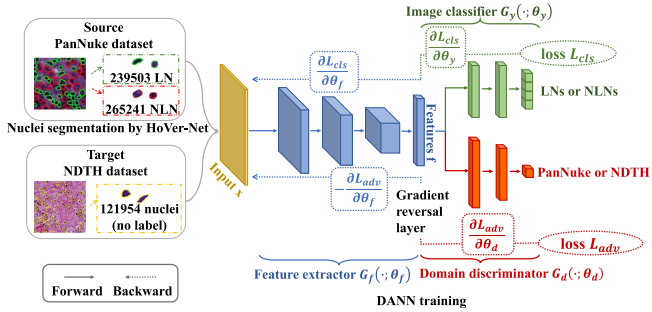

Fig. 3. Domain adversarial neural network (DANN) training strategy based on the annotated lymphocyte nuclei dataset from PanNuke and our collected nuclei dataset segmented by the pretrained HoVer-Net model. LN and NLN denote lymphocyte nuclei and non-lymphocyte nuclei, respectively.

map as the fine-scale feature on each $d \times d$ tile image based on the previous predictions. Meanwhile, i order to learn the generic global context information, we get the coarse-scale feature from the original image by $d \times d$ mean pooling operation to make it the same size as the lymphocyte density map. Additionally, we utilize a reverse operation on the lymphocyte density map to make it be an attention map, termed as the lymphocyte density attention (LDA), which is compatible with the original H&E image in terms of the intensity distribution. Then, a cross-scale attention guidance mechanism based on a U-shape backbone with four-channel inputs is proposed to process the above cross-scale images. It learns the macroscopic features from the coarse-scale H&E image and microscopic features from the fine-scale lymphocyte density attention. Additionally, the proposed segmentation model is trained with bounding box annotations, which is weakly supervised for the segmentation task. Furthermore, a noise-sensitive constraint with a signed distance function loss (SDF) is used in the training procedure to explicitly learn the TLS distribution and avoid huge performance drops caused by tiny predicted errors.

A. Segmentation and Classification for Lymphocyte Nuclei

The immune response in pancreatic tumors is usually not strong, and the distribution of immune cells is relatively sparse. In contrast, we note that the lymphocyte nuclei' morphological characteristics are independent of the organs in which they are located. The lymphocyte nuclei on pathology images of other organs should be morphologically identical to the pancreatic lymphocyte nuclei, although these images have different backgrounds. Therefore, we considered transferring the lymphocyte nuclei from a public dataset with lymphocyte nuclei annotations of other organs to our pancreatic pathological dataset without lymphocyte nuclei annotations.

To this end, we introduce the combination of a pretained baseline model for nuclei segmentation and a domain adversarial network for lymphocyte nuclei classification shown in Fig. 2. HoVer-Net is a robust baseline model for nuclei segmentation and has demonstrated state-of-the-art performance on various nuclei segmentation tasks [14]. There are three branches in the HoVer-Net architecture: the segmentation branch for getting coarse prediction, the structure branch for solving the nuclei overlapping, and the classification branch for nuclei recognition. We use the segmentation and structure branches from the pretrained HoVer-Net in the PanNuke dataset [15] to segment the nuclei for our task. As for the classification, we train a domain adversarial neural network (DANN) [16] to get better performance on lymphocyte nuclei recognition. A domain adversarial training strategy is used in the classification network. Fig. 3 demonstrates our training strategy for domain adversarial learning. We crop 239503 lymphocyte nuclei images and 265241 non-lymphocyte nuclei images from the PanNuke dataset with nuclei annotations. Then we obtain 121954 nuclei images with unknown categories (lymphocyte or non-lymphocyte nuclei) from the collected dataset based on the segmentation results from the pretrained HoVer-Net. We choose the former as the source domain and the latter as the target domain to train the domain adversarial network for lymphocyte nuclei recognition. ResNet18 [17] is used as the feature extractor, and two loss terms are calculated as follows.

$$L_{cls} = \sum_{i=1}^{M} y_i \log \frac{1}{G_y(G_f(x_i))}, \tag{1}$$

$$L_{adv} = \sum_{i=1}^{M} d_i \log \frac{1}{G_d(G_f(x_i))} + (1 - d_i) \log \frac{1}{G_d(G_f(x_i))}, \tag{2}$$

where G_f, G_y, and G_d denote the feature extractor, image classifier, and domain discriminator, respectively. M denotes the number of training samples. x_i, y_i, and d_i denote the image input, nucleus category annotation, and domain label of sample i, respectively. The overall loss of DANN is defined by

$$L_{DANN} = L_{cls} + L_{adv} \tag{3}$$

B. Cross-Scale Attention Guidance Mechanism

TLSs are the particular forms of lymphocyte aggregations, and it is natural to look for TLSs where lymphocytes are relatively aggregated. Therefore, we calculate the lymphocyte density map based on the detected lymphocyte nuclei and convert it to be attention map to guide the detection of TLSs. To obtain the

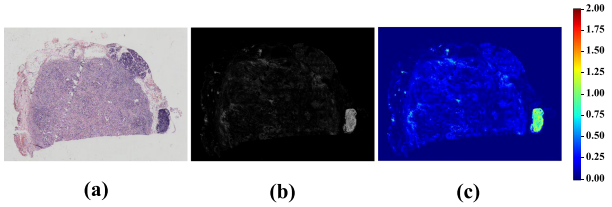

Fig. 4. Visualization of the calculated lymphocyte density map of a WSI from our collected dataset. (a) WSI; (b) Gray-scale image of the calculated lymphocyte nuclei density map, high density regions are denoted by high grayscale; (c) Heatmap of the calculated lymphocyte nuclei density. The number on the colorbar denotes the lymphocyte nuclei density calculated as the number of lymphocyte nuclei within a $25 \times 25 \mu m^2$ region.

lymphocyte density map on a WSI, we count the number N_{ij} of predicted lymphocyte nuclei in each non-overlapping $d \times d$ patch at location (i, j). Each pixel occupies 25×25 μm^2. Then we get a gray-scale lymphocyte density map by the following formulation:

$$D_{ij} = 255 \times \frac{N_{ij} - N_{\min}}{N_{\max} - N_{\min}}, \tag{4}$$

where $N_{\max}$ and $N_{\min}$ are the maximum and minimum of the predicted number of lymphocyte nuclei among the patches. Fig. 4 shows the lymphocyte density map calculated for a WSI from our collected dataset. The channel-wise attention mechanism has demonstrated satisfactory performance for medical image analysis [18]. The knowledge-based attention map can effectively guide the network to focus on the target regions. We split the three channels of the original H&E images into blue, green, and red channels shown in Fig. 1. It can be found that the target TLS regions show low intensity in each color channel. We add a reverse operation to make the predicted lymphocyte density map compatible with the original H&E image to generate our desired attention map. Each lymphocyte density attention A_{ij} at location (i, j) is calculated as $A_{ij} = 255 - D_{ij}$. Then a cross-scale attention guidance network is established on the U-shape backbone architecture with skip connections. There are four channel inputs of the backbone to jointly learn the coarse-scale features from the original H&E images and the fine-scale features from the calculated lymphocyte density attention.

C. SDF Loss for Noise-Sensitive Constraint

Considering that accurate positioning usually has high priority than finding the boundary of a TLS for clinical issues. Therefore, we train the proposed network with the bounding box annotations from the experts, which is weakly supervised for a segmentation task. Generally, it is difficult for a semantic segmentation network to detect small objects by training with general segmentation loss, such as Dice loss and cross-entropy (CE) loss, because these small regions almost have no impact on the overall loss. However, minor prediction errors can lead to misdiagnosis. To solve this problem, we introduce a noise-sensitive constraint by embedding a signed distance function (SDF) loss into the overall loss function for the training procedure. The combination of SDF and deep learning was firstly designed for high quality shape representation, interpolation and

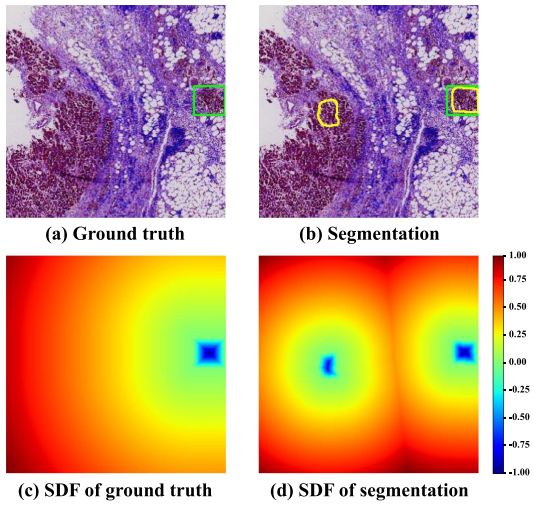

Fig. 5. Visual comparison between the SDF of ground truth and SDF of segmentation. (a) and (b) are the ground truth (green rectangle) and the segmentation (yellow contours) of TLSs. (c) and (d) are the corresponding signed distance function representation. The number on the colorbar denotes the distance to the mask boundary. It can be found that tiny segmentation errors can significantly affect the SDF distribution.

completion from partial and noisy 3D input data [19]. Recently, the SDF loss has been adopted for medical image segmentation and demonstrated the advantage in reducing the small noise predictions [20]. The signed distance function of a predicted binary mask Y can be calculated as follows:

$$SDF(x, Y) = \begin{cases} \min\limits_{y \in \partial Y} -d(x, y), & x \in Y_{in} \\ 0, & x \in \partial Y \\ \min\limits_{y \in \partial Y} d(x, y), & x \in Y_{out}, \end{cases} \tag{5}$$

where $d(x, y)$ is the Euclidian distance between x and y, and Y_{in}, Y_{out} and ∂Y denote the inside, the outside and the boundary of the object, respectively. Then we get the signed distance loss by calculating the mean square error (MSE) between the SDFs of segmentation and ground truth:

$$L_{SDF}(Y, \hat{Y}) = \frac{\sum_{i=1}^{N} \left(SDF(x_i, Y) - SDF\left(x_i, \hat{Y}\right) \right)^2}{N}. \tag{6}$$

where N is the number of pixels in an input image. As shown in Fig. 5, small segmentation errors significantly impact the SDF distribution. Besides, the traditional Dice and CE losses are also used in the training procedure. They can be calculated as follows:

$$L_{Dice} = -\frac{2 \sum_{i=1}^{N} s_i g_i}{\sum_{i=1}^{N} s_i^2 + \sum_{i=1}^{N} g_i^2}, \tag{7}$$

$$L_{CE} = -\frac{\sum_{i=1}^{N} g_i ln p_i}{N}, \tag{8}$$

where s_i and g_i denote the predicted segmentation and the ground truth of pixel i, respectively. p_i denotes the softmax output of s_i. Therefore, the overall loss function of our proposed TLSs segmentation network can be formulated as follows:

$$L = L_{Dice} + L_{CE} + wL_{SDF}, \tag{9}$$

where $w > 0$ is a weight to achieve a trade-off between different loss terms.

III. Experiments

A. Dataset

In this work, we evaluate our proposed method on two datasets collected from Nanjing Drum Tower Hospital (NDTH) and Jiangsu Province Hospital of Chinese Medicine (JHCM), respectively. The NDTH dataset consists of 38 WSIs from 12 surgical pathology-confirmed PDAC patients. The JHCM dataset is composed of 57 WSIs from 41 PDAC patients. To verify the generalization and the performance in few-shot learning of our proposed method, we train and validate our model on the smaller-scale NDTH dataset and test on the larger-scale JHCM dataset. This study was approved by the Ethics Committee of Nanjing Drum Tower Hospital and Jiangsu Province Hospital of Chinese Medicine.

The TLSs on the WSIs are annotated with bounding boxes by two experienced pathologists. The NDTH dataset is randomly divided into five folds. The numbers of the cases belonging to each fold are 8, 8, 8, 7, 7. All the experiments on the NDTH dataset are based on the five-fold cross-validation. The JHCM dataset is only used for testing the ensemble models trained on the NDTH dataset. (i.e., zero-shot evaluation).

B. Evaluation Metrics

We employ the precision, recall and F_β score to measure the detection accuracy in our experiments, which can be calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{12}$$

where TP, FP and FN denote the true positive, false positive and false negative TLS predictions, respectively which are calculated based on the intersection over union (IoU) between the connected components of the predictions and ground truth masks. We set the threshold of the IoU as 10% in the experiments. Generally, F_1 score (i.e., $\beta = 1$ in (15)) is used to be the harmonic mean of both the precision and recall. However, because the false negative predictions generally bring greater harm than the false positive predictions in clinical issues, which means the recall metric is more important than the precision metric. Therefore, we also use the F_2 score (i.e., $\beta = 2$ in (15)) to evaluate the detection performance with greater weight on recall.

Because the TLS is an aggregation form of lymphocyte nuclei, it is common for experts to have different annotations due to their different experiences. Fig. 6 shows two examples to describe the two experts' annotation difference. It is difficult to have a consistent guideline for different experts to judge whether there is one or two TLSs in an area with irregular boundary during the time-consuming TLS labeling. Therefore, it is inevitable for experts to have annotation bias when labeling. The elimination of bias often requires the participation of more authoritative experts for revisions, which is usually more labor-intensive and time-consuming. To address this problem, we introduce a new metric to reflect the TLS detection performance evaluated based on the annotations without bias in the local areas with irregular boundaries.

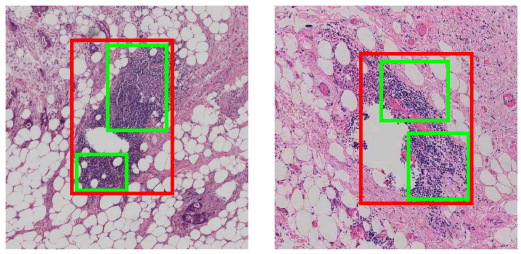

Fig. 6. Two examples to illustrate the annotation difference by different experts. The TLS annotations by two experts which are denoted by red and green rectangles, respectively.

Without loss of generality, since the precision describes how many positive cases are true among all the model predictions, we introduce the TPS here instead of TP to represent the number of the connected components in segmentation that overlap with ground truths. Similarly, the recall describes how many annotated boxes are truly predicted, we introduce the TPB instead of TP to represent the number of annotated bounding boxes that overlap with the segmentation results. It can be seen that the TPS and TPB are generalizations of the general TP without the one-on-one constraint. Then, we define the corresponding segmentation precision (SP) and box recall (BR) as follows:

$$SP = \frac{TPS}{TPS + FP}, \tag{13}$$

$$BR = \frac{TPB}{TPB + FN}. \tag{14}$$

Similarly, the general F_β score (GF_β) can be calculated as follows:

$$GF_\beta = (1 + \beta^2) \times \frac{SP \times BR}{\beta^2 \times SP + BR}. \tag{15}$$

Fig. 7 shows three examples to illustrate the calculation of the proposed metric.

C. Implementation Details

We use the 2D nnU-Net as the network backbone. Images are normalized to the range of 0–1. Specifically, all foreground voxels in the training set are collected, and an automated level-window-like clipping of intensity values is performed based on the 0.5 and 99.5th percentile of these values. The data is then normalized with the global foreground mean and standard deviation. All spacings within the training data are collected and for each axis the median is chosen as the target spacing. All training cases are then resampled with third order spline interpolation. The training patches are randomly cropped with size of 320×320 from the multi-channel image inputs. Additionally, we set the

TABLE I

ABLATION STUDY RESULTS OF THE PROPOSED METHOD ON TWO COLLECTED DATASETS. EXPERIMENTAL RESULTS ON THE NDTH DATASET ARE BASED ON THE FIVE-FOLD CROSS-VALIDATION. THE ENSEMBLE MODEL IS USED FOR TESTING ON THE JHCM DATASET. W/O MEANS REMOVING THE CORRESPONDING MODULE

Datasets	Methods	P	R	F1	F2	SP	BR	GF1	GF2
NDTH	Proposed	**80.0**	**83.6**	**80.9**	**82.3**	**81.6**	**85.6**	**82.9**	**84.3**
	Proposed w/o L_{SDF}	78.1	81.6	79.0	80.4	79.2	83.3	80.5	82.0
	Proposed w/o LDA	73.7^{+}	83.0	76.9^{+}	80.1	74.8^{*}	84.9	78.3^{*}	81.7^{+}
	Proposed w/o LDA, L_{SDF}	75.3	79.5^{+}	76.3^{*}	77.9^{+}	76.5^{+}	80.3^{+}	77.2^{*}	78.7^{*}
JHCM	Proposed	**74.7**	86.5	**77.8**	**82.2**	**80.9**	88.0	**82.9**	**85.6**
	Proposed w/o L_{SDF}	70.4	**87.5**	75.9	81.7	76.5^{*}	**89.0**	80.6^{+}	84.9
	Proposed w/o LDA	68.9^{+}	78.3^{*}	71.5^{*}	74.9^{*}	74.6^{*}	79.5^{*}	74.6^{*}	76.5^{*}
	Proposed w/o LDA, L_{SDF}	73.0	74.0^{*}	71.8^{*}	72.7^{*}	76.7^{+}	75.3^{*}	73.8^{*}	73.8^{*}

P, R, F1 and F2 denote the precision, recall, F_1 score and F_2 score, respectively. SP, BR, GF1 and GF2 represent the introduced segmentation precision, box recall, GF_1 score and GF_2 score, respectively. $^{+}$ and * denote the statistically significant level $p < 0.05$ and $p < 0.01$, respectively, when compared with the proposed method.

The bold values stand for the best performance in the comparison experiments.

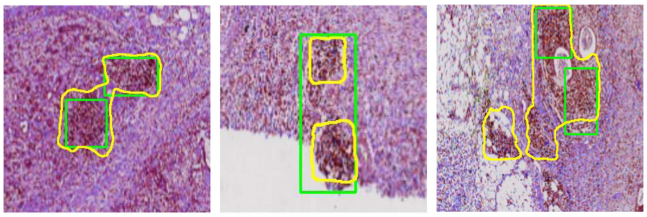

Fig. 7. Three examples to illustrate the proposed evaluation metrics.(a) When one predicted segmentation denoted by yellow contour covers two annotated boxes denoted by green rectangles, there are $TPS = 1$ and $TPB = 2$. (b) When one annotated box denoted by green rectangle is covered by two predicted segmentation denoted by yellow contours, there are $TPS = 2$ and $TPB = 1$. (c) When the segmentation mask and the bounding box partially overlap with each other, the metric calculation is base on the IoU between the connected components of the predictions and ground truth masks. In this case, there are $TPS = 1$ and $TPB = 2$.

patch size d for coarse-scale pooling and fine-scale density calculation as 25 μm based on medical experts' experience.

D. Ablation Study

In this section, we present the ablation study to verify the effectiveness of our proposed LDA and SDF in our method. We show the ablation study results in Table I to investigate the individual impact of the proposed LDA and SDF module, the best performance is shown in bold.

1) Effectiveness of LDA: Without using LDA, the performance significantly drops on our collected datasets, leading to a decrease of 5.8%–6.3% in precision, 0.6%–8.2% in the recall, 4.0%–6.3% in F_1 score, and 2.2%–7.3% in F_2 score. For the introduced metrics, there is a performance decrease of 6.3%–6.8% in SP, 0.7%–8.5% in BR, 4.6%–8.3% in GF_1 score, and 2.6%–9.1% in GF_2 score without the LDA. Fig. 8 shows the visual comparison of the ablation study for the proposed LDA. It can be found that the false positive predictions with low attention and false negative predictions with high attention can be efficiently captured.

2) Effectiveness of SDF: Without using the SDF loss, the performance also drops on our collected datasets, leading to a decrease of 1.9%–4.3% in precision, 1.9% in F_1 score, and 0.5%–1.9% in F_2 score. For the introduced metrics, there is a performance decrease of 2.4%–4.4% in SP, 2.3%–2.4% in GF_1 score, and 0.7%–2.3% in GF_2 score without the SDF loss. It should be pointed that the SDF loss leads to a improvement in recall and BR on the NDTH dataset while it leads to a decrease in recall and BR on the JHCM dataset, which indicates that the SDF loss improves the overall performance of the detection model by mainly reducing the tiny false positive predictions instead of false negative predictions with relatively large sizes. Fig. 9 presents the visual comparison of the ablation study for

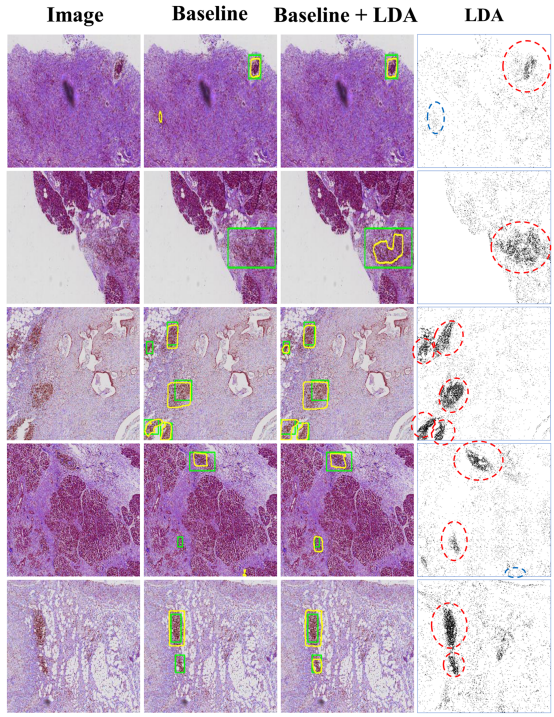

Fig. 8. Visual comparison of the ablation study for the proposed LDA. The baseline model is the proposed method without LDA and SDF loss. Regions with high and low attention on the calculated LDAs are signed with red and blue dashed circles, respectively. It can be found that the calculated LDA can significantly improve the TLS detection performance.

TABLE II
QUANTITATIVE COMPARISON RESULTS FOR TLSS DETECTION BY THE PROPOSED METHOD WITH DIFFERENT ATTENTION INPUT ON TWO COLLECTED DATASET. EXPERIMENTAL RESULTS ON THE NDTH VALIDATION SET AND THE JHCM DATASET ARE PRESENTED

Datasets	Attention	P	R	F1	F2	SP	BR	GF1	GF2
NDTH	Coarse density map	71.4*	82.0	75.2*	78.8^{+}	74.2^{+}	83.1	77.3*	80.4*
	LDA	**80.0**	**83.6**	**80.9**	**82.3**	**81.6**	**85.6**	**82.9**	**84.3**
JHCM	Coarse density map	69.1^{+}	83.7	73.9	78.8	73.8*	85.6	77.9*	81.8*
	LDA	**74.7**	**86.5**	**77.8**	**82.2**	**80.9**	**88.0**	**82.9**	**85.6**

P, R, F1 and F2 denote the precision, recall, F_1 score and F_2 score, respectively. SP, BR, GF1 and GF2 represent the introduced segmentation precision, box recall, GF_1 score and GF_2 score, respectively. $^{+}$ and * denote the statistically significant level $p < 0.05$ and $p < 0.01$, respectively.
The bold values stand for the best performance in the comparison experiments.

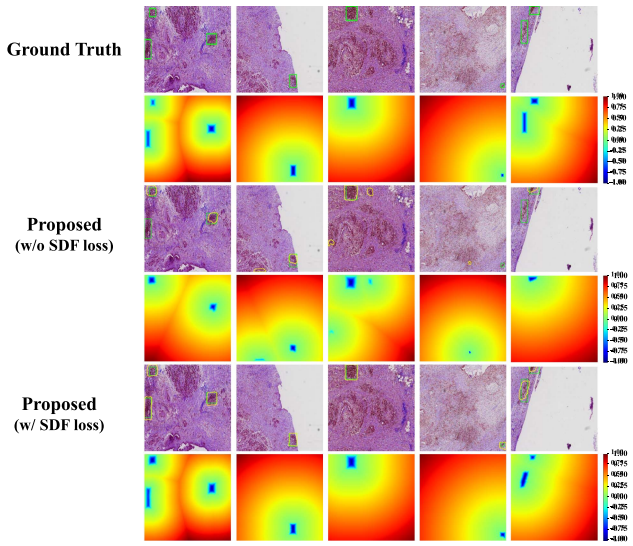

Fig. 9. Visual comparison of the ablation study for the SDF loss. Segmentation and ground truth are denoted by green and yellow contours in pathological images, respectively. The SDF maps are shown under the corresponding images.

the SDF loss. We can see that tiny segmentation errors can cause huge losses in terms of the signed distance function map, and the SDF loss can significantly reduce the tiny false positive predictions.

3) Effectiveness of LDA and SDF: Additionally, it can be observed that without using LDA and SDF losses leads to the worst performance in the recall, F_1 score, F_2 score BR, GF_1 score, GF_2 score, indicating that both the proposed LDA and SDF losses are essential for performance improvements. It should be pointed that the baseline model (the proposed method without LDA and SDF loss) achieves the best performance in segmentation precision on NDTH dataset, while the proposed method outperforms the baseline model by 0.29% in the precision and achieves the best performance in all other metrics.1.7%–4.7% in precision, 4.1%–12.5% in recall, 4.6%–6.0% in F_1 score, 4.4%–9.5% in F_2 score, 4.2%–5.1% in SP, 5.3%–12.7% in BR, 5.7%–9.1% in GF_1, 5.6%–11.8% in GF_2.

4) Comparison With the Coarse Density Map: We add the reverse operation to make the predicted lymphocyte density map compatible with the original H&E image to generate our desired attention map. In order to study the performance difference between the coarse density map and the designed LDA, we conduct extra comparison experiments as follows. We use the coarse density map as the fourth channel input instead of the designed LDA on the JHCM dataset and compare the TLS detection performance between them. Statistical testing is also provided. Table II shows the experimental results. It can be observed that the designed LDA has significant superiority than the coarse density map.

E. Comparison With Other State-of-the-art Segmentation Methods

In this section, we compare our proposed method with other state-of-the-art (SOTA) methods for TLS detection. The UNet [21], Deeplab v3+[22], and nnUNet [23] are included in our experiments for their excellent performance in medical image segmentation. All these methods are trained on the NDTH training set and evaluated on the NDTH validation set and the JHCM dataset. Quantitative comparison results are presented on Table III.

Experimental results in Table III demonstrate that our proposed method significantly outperforms these SOTA methods, and achieves the best performance in almost every evaluation metric. We observe that the generic UNet [21] and deeplab v3+ [22] always have conservative predictions by minimizing the false positive predictions as much as possible, despite causing many false negative predictions. Therefore, they have a higher precision and SP than ours, but leading to 15.2%–26.7%, 7.0%–12.8%, 12.0%–19.4%, 16.3%–32.0%, 7.1%–19.6% and 12.7%–26.0% decrease in recall, F_1 score, F_2 score, GF_1 score and GF_2 score, respectively comparing with ours. Fig. 10 shows the visual comparison of TLS detection results between the above methods and ours.

F. Experimental Results for the Nuclei Segmentation and Classification

Actually, the results of the previous nuclei detection stage may affect the results of the next tertiary lymphoid structure (TLS) detection stage in general. We term the designed combination of the baseline HoVer-Net for nuclei segmentation and the domain adversarial neural network (DANN) for lymphocyte nuclei classification as Transferred HoVer-Net with DANN (THD). To study the influence of the potential bad segmentation from the designed THD, we add extra studies as follows.

Firstly, we show the nuclei segmentation and classification results predicted by the designed THD. The experts meticulously annotated the lymphocyte and non-lymphocyte nuclei pixel by

TABLE III
QUANTITATIVE COMPARISON RESULTS WITH OTHER SOTA METHODS FOR TLSs DETECTION ON TWO COLLECTED DATASETS. EXPERIMENTAL RESULTS ON THE NDTH DATASET ARE BASED ON THE FIVE-FOLD CROSS-VALIDATION. THE ENSEMBLE MODEL IS USED FOR TESTING ON THE JHCM DATASET

Datasets	Methods	P	R	F1	F2	SP	BR	GF1	GF2
NDTH	UNet [21]	**83.1**	68.4*	73.9$^+$	70.3*	**86.2**	69.3*	75.8$^+$	71.6*
	DeepLab v3+ [22]	81.0	67.4*	71.9*	68.9*	84.7	68.5*	74.7$^+$	70.6*
	nnUNet [23]	75.3	79.5$^+$	76.3*	77.9$^+$	76.5$^+$	80.3$^+$	77.2*	78.7*
	Proposed	80.0	**83.6**	**80.9**	**82.3**	81.6	**85.6**	**82.9**	**84.3**
JHCM	UNet [21]	71.8	63.4*	67.6*	65.6*	75.8$^+$	65.6*	70.8*	68.1*
	DeepLab v3+ [22]	72.8	59.8*	65.0*	62.8*	71.4$^+$	56.0*	63.3*	59.6*
	nnUNet [23]	73.0	74.0*	71.8*	72.7*	76.7$^+$	75.3*	73.8*	73.8*
	Proposed	**74.7**	**86.5**	**77.8**	**82.2**	**80.9**	**88.0**	**82.9**	**85.6**

P, R, F1 and F2 denote the precision, recall, F_1 score and F_2 score, respectively. SP, BR, GF1 and GF2 represent the introduced segmentation precision, box recall, GF_1 score and GF_2 score, respectively. $^+$ and * denote the statistically significant level $p < 0.05$ and $p < 0.01$, respectively. respectively, when compared with the proposed method.
The bold values stand for the best performance in the comparison experiments.

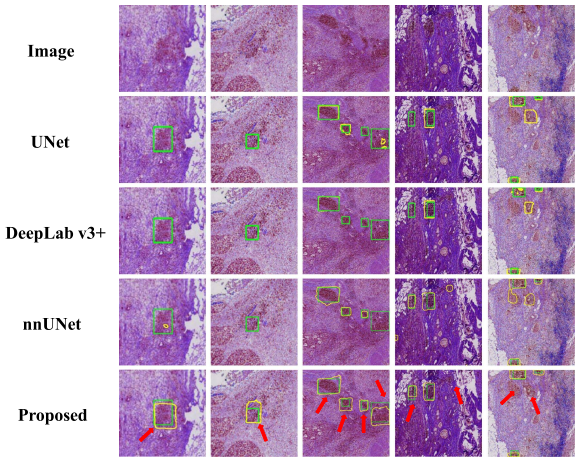

Fig. 10. Visual comparison results between the SOTA segmentation methods and our proposed method. The green and yellow denote the ground truth and predictions, respectively. Improved predictions are signed by red arrows.

pixel on 50 cropped patches on the JHCM dataset, and we termed the labeled dataset as the reference dataset. We test the designed THD model for lymphocyte nuclei segmentation and classification on the reference dataset and get the traditional precision, recall, and F1 score metrics. We used the pre-trained HoVer-Net with the classification branch on the PanNuke dataset as the baseline model for comparison. Because the pre-trained HoVer-Net on the PanNuke dataset is used to recognize six types of nuclei, including the lymphocyte nuclei, we classify the other five types of nuclei as non-lymphocyte nuclei. The experimental results are presented in Table IV . We can observe that although the baseline model achieves better precision, the designed THD shows significant superiority in the recall and F1-score metrics, which demonstrate the effectiveness of the embedding DANN. Additionally, we analyzed that the baseline model can get slightly better precision may be due to the conservative predictions, and it is difficult for the baseline model to distinguish the other nuclei with similar characteristics to the lymphatic nuclei, such as the neoplastic nuclei. We present the visual comparison in Fig. 11.

TABLE IV
SEGMENTATION AND CLASSIFICATION RESULTS FOR LYMPHOCYTE NUCLEUS ON THE REFERENCE DATASET

Methods	P	R	F1
Pretrained HoVer-Net	**96.2**	61.8*	72.5$^+$
Tranferred HoVer-Net with DANN (**Proposed**)	93.5	**86.0**	**88.8**

$^+$ and * denote the statistically significant level $p < 0.05$ and $p < 0.01$, respectively.
The bold values stand for the best performance in the comparison experiments.

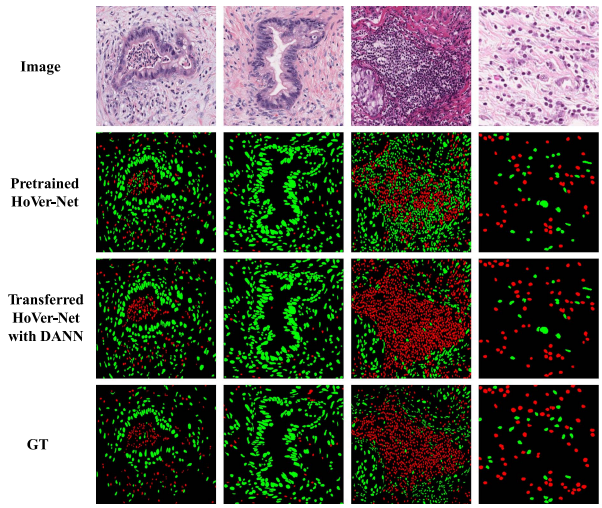

Fig. 11. Visual comparison between the proposed transferred HoVer-Net with the HoVer-Net pretrained on the PanNuke datatset for lymphocyte nuclei segmentation and classification. The red and green mask denote the lymphocyte and non-lymphocyte nuclei, respectively.

Then, to study whether the designed THD model in the first stage has achieved satisfactory lymphocyte nuclei detection performance, we use the ground truth of the lymphocyte nuclei to replace the predicted results by the THD model. The ground truth is obtained by means of that two experts were asked to revise the segmentation and classification results of lymphocyte nuclei on each WSI of the JHCM dataset. In order to improve the labeling efficiency, all the revisions are based on the point

TABLE V
THE EXPERIMENTAL RESULTS OF THE PROPSOED MODEL FOR TLSS DETECTION WITH DIFFERENT ATTENTION INPUTS

Datasets	LDA version	P	R	F1	F2	SP	BR	GF1	GF2
JHCM	Prediction	**74.7**	86.5	**77.8**	82.2	**80.9**	88.0	**82.9**	85.6
	Ground Truth	73.1	**87.5**	77.3	**82.5**	79.3	**89.3**	82.2	**85.8**
	p-value	0.493	0.644	0.794	0.906	0.105	0.261	0.343	0.677

P, R, F1 and F2 denote the precision, recall, F_1 score and F_2 score, respectively. SP, BR, GF1 and GF2 represent the introduced segmentation precision, box recall, GF_1 score and GF_2 score, respectively. The P-value is provided for statistical testing.
The bold values stand for the best performance in the comparison experiments.

TABLE VI
QUANTITATIVE COMPARISON RESULTS FOR TLSS DETECTION BY THE PROPOSED METHOD WITH DIFFERENT SDF LOSS WEIGHT ON TWO COLLECTED DATASET. EXPERIMENTAL RESULTS ON THE NDTH VALIDATION SET AND THE JHCM DATASET ARE PRESENTED

Datasets	SDF weight	P	R	F1	F2	SP	BR	GF1	GF2
NDTH	$\omega = 0$	78.1	81.6	79.0	80.4	79.2	83.3	80.5	82.0
	$\omega = 0.5$	**81.6**	79.5	80.1	79.6	83.3	82.8	82.6	82.6
	$\omega = 1$	80.0	**83.6**	**80.9**	**82.3**	81.6	**85.6**	**82.9**	**84.3**
	$\omega = 1.5$	80.1	82.1	80.5	81.3	81.0	84.2	82.0	83.1
	$\omega = 2$	77.1	82.5	79.2	81.0	78.4	84.0	80.6	82.5
JHCM	$\omega = 0$	70.4	87.5	75.9	81.7	76.5	89.0	80.6	84.9
	$\omega = 0.5$	**75.6**	85.5	**78.4**	81.9	**82.5**	87.0	**83.1**	85.0
	$\omega = 1$	74.7	86.5	77.8	82.2	80.9	88.0	82.9	85.6
	$\omega = 1.5$	71.3	**88.1**	76.8	**82.6**	77.1	**89.9**	81.3	**85.8**
	$\omega = 2$	73.1	87.0	77.1	82.1	78.4	88.2	81.0	84.6

P, R, F1 and F2 denote the precision, recall, F_1 score and F_2 score, respectively. SP, BR, GF1 and GF2 represent the introduced segmentation precision, box recall, GF_1 score and GF_2 score, respectively.
The bold values stand for the best performance in the comparison experiments.

annotations, which can be easily conducted and effective to calculate the density maps. The LDA maps are generated by these two different lymphocyte nuclei density maps. We compare the TLS detection performance of the proposed method with these two versions of LDA and conduct statistical testing experiments. The experimental results are shown in Table V. It can be observed that the LDA map generated by the ground truth can only slightly improve the TLS detection performance without statistical significance. Therefore, we verify that the designed THD model can achieve satisfactory performance for lymphocyte nuclei detection in our experiment.

Besides, we analyze the following possible reasons for the above results: (1) The proposed method is with a four-channel encoder architecture. In addition to the LDA map generated by the lymphocyte nuclei detection results, the original RGB image channels also have great learning weights for the TLS feature extraction. (2) Because the TLS is an aggregation form of lymphocyte nuclei, the predicted LDA maps generated by these imperfect lymphocyte nuclei detection results can still provide rich structure information and feature cues to improve the TLS detection when the lymphocyte nuclei detection achieves satisfactory performance. (3) The proposed method is learned with the LDA maps generated by these imperfect lymphocyte nuclei detection results, so the whole network tends to have a certain ability for fault tolerance.

G. The Contribution of SDF Weight to the Overall Loss

In our experiments, we adopt the default weight for training the domain adversarial neural network and set the weight of each term as 1. We adopted the same weight strategy on the overall loss of the proposed method. To study how the SDF's weighted contribution influences the overall loss, we conduct the extra experiments as follows.

We set the SDF weight as 0, 0.5, 1, 1.5, 2 in sequence, and trained five different TLS detection models. The experiments on the NDTH dataset are based on the five-fold cross-validation and the JHCM dataset is used for testing the ensemble model trained on the NDTH dataset. Statistical testing is also provided. Table VI shows the experimental results, and the best second-best results are shown in bold and red, respectively. It can be observed that the proposed method achieves the most balanced and stable lead performance when the SDF loss weight is set to 1.

H. Insight of the Proposed Metrics

In order to study whether the proposed metrics can reflect the TLS performance without annotation bias. We invite three experts to jointly revise the TLS annotations case by case on the NDTH dataset and recalculate the traditional metrics based on these annotations. We draw the line chart of the performance comparison between the original metrics with bias, the introduced metrics, and the original metrics without bias in Fig. 12. Our introduced segmentation precision (SP), box recall (BR), general F1 score (GF1), and general F2 score (GF2) correspond to the original precision (P), recall (R), F1 score (F1), and F2 score (F2), respectively. It can be observed that the TLS detection performance evaluated by our introduced metrics is closer to the performance evaluated by the annotations without bias in recall and F2 score. We give a more striking example presented in Fig. 13. It should be pointed out that our introduced metrics

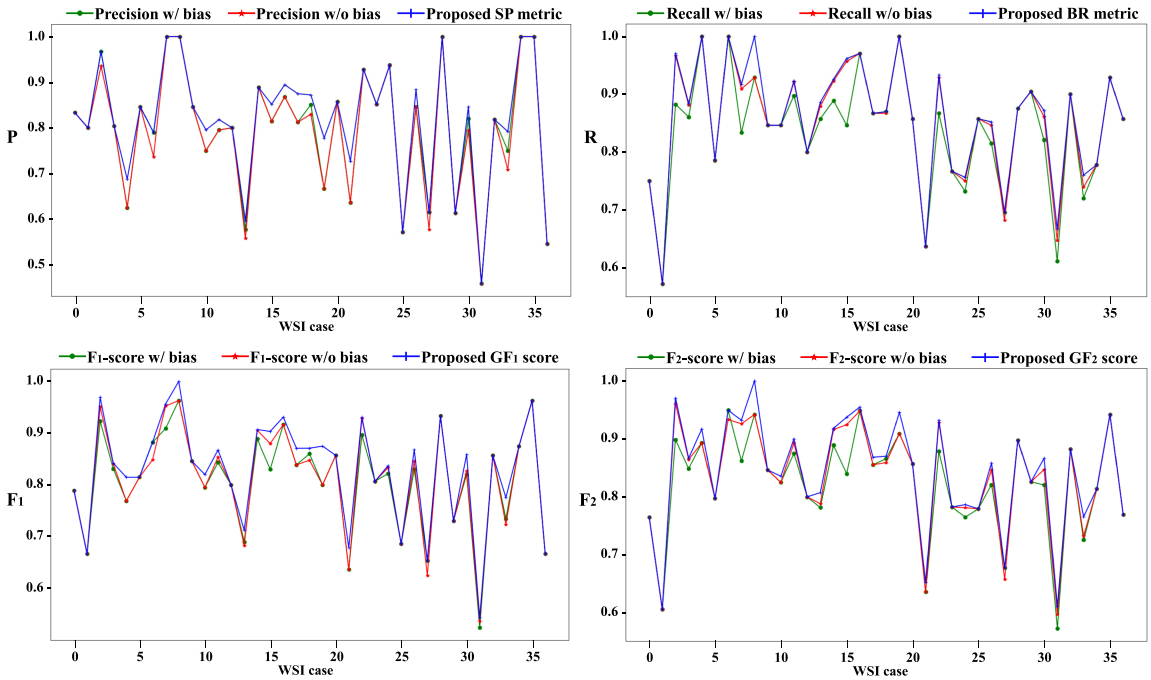

Fig. 12. Line charts of TLS detection performance comparison between the original metrics with bias (shown in green), the introduced metrics (shown in blue) and the original metrics without bias (shown in red).

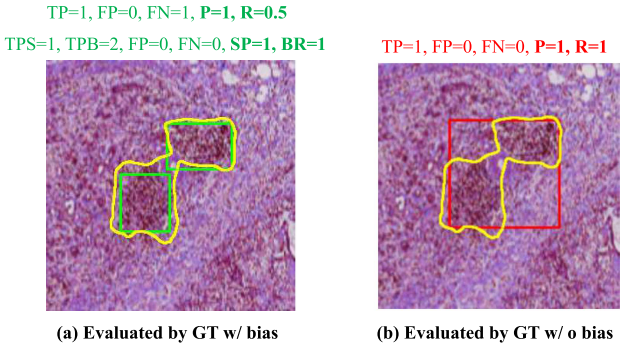

Fig. 13. Examples to illustrate the advantage of the introduced evaluation metrics in reflecting the TLS detection performance without annotation bias. (a) The original metrics and the introduced metrics evaluated by the annotations with bias. (b) The original metrics evaluated by the annotations without bias.

are not just sensitive to the recall or the F2 score metrics. The difference between the performance evaluated by the original metrics and the introduced metrics reflects the annotation bias on the experimental dataset. More specifically, assume that the green and red annotations in Fig. 6 denote the TLS annotations with and without bias, respectively. If the situation in the first row commonly occurs on the experimental dataset, then the introduced metrics can reflect the recall better than the original metrics evaluated by the annotations with bias. On the contrary, if the situation in the second row commonly occurs on the experimental dataset, then the introduced metrics can reflect the precision better than the original metrics evaluated by the annotations with bias.

In conclusion, the introduced metrics can reflect the TLS performance evaluated by the annotations without bias in the recall or precision metrics better than the original metrics. Although the order of the model performance may not change, the difference between the values of the original and introduced metrics reflects the bias in labeling, which can be easily calculated instead of making revisions by more authoritative experts.

I. Application for Studying the Relationship Between TLS Density and Peripancreatic Vascular Invasion

For patients with pancreatic tumors accompanied by peripancreatic vascular invasion, especially venous invasion, whether the peripancreatic vessel is invaded determines the direction of treatment, and also affect the survival time and prognosis of patients [24], [25], [26]. In this section, we apply our proposed TLS detection method to study the relationship between TLS density and peripancreatic vascular invasion.

We collect the clinical information from 12 patients in NDTH dataset and 40 patients (there is a patient whose clinical information is not available) in JHCM datatset to carry out our statistical experiments. There are 19 patients without peripancreatic vascular invasion and 33 patients with peripancreatic vascular invasion. Therefore, we divide the 52 samples into two groups: the no-invasion group and the invasion group. Then we apply our proposed method to detect the TLSs on their WSIs, and calculate the TLS density of the patients in the two groups, respectively. The TLS density is calculated as the ratio of the TLS number to the WSI's area.

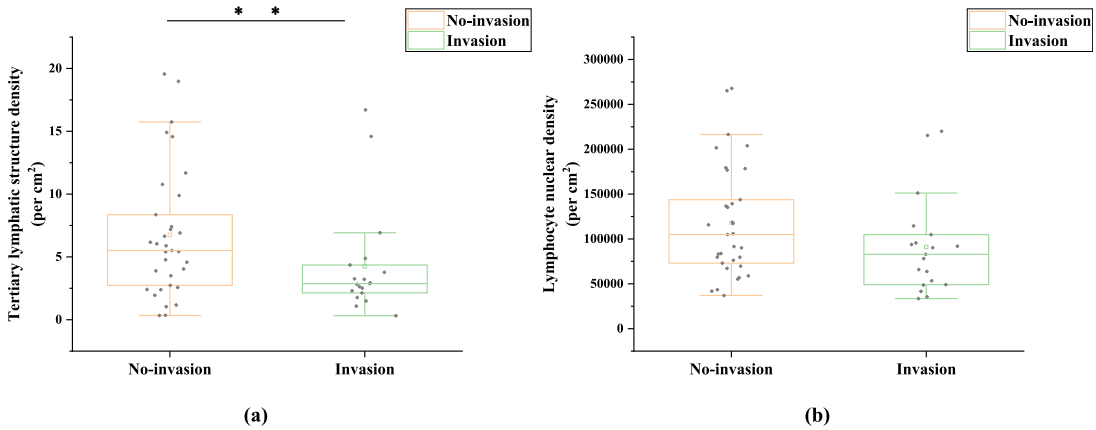

Fig. 14. Distribution of (a) TLS density and (b) lymphocyte density for two groups (the no-invasion group and the invasion group). $**$ indicates $p < 0.05$ using the Mann-Whitney U test.

The Shapiro-Wilk (S-W) test [27] is conducted to check whether the TLS densities of the 52 patients obey the normal distribution. Based on the S-W test results of $p < 0.05$ and $p < 0.001$ on the invasion and no-invasion group, respectively (i.e., they do not obey the normal distribution), we use the Mann-Whitney U test [28] to observe the correlation between the TLS density and the peripancreatic vascular invasion. We get the result of $p = 0.03$ for the Mann-Whitney U test, which indicates that the TLS density is indeed related to the peripancreatic vascular invasion.

For comparison, we also study on the correlation between the lymphocyte density and the peripancreatic vascular invasion. The same statistical testing experiments are conducted for the lymphocyte density. We get the S-W test results of $p < 0.01$ on the invasion group (i.e., they do not obey the normal distribution). We use the Mann-Whitney U test to observe the correlation between the TLS density and the peripancreatic vascular invasion. The we get the result of $p = 0.1$ for the Mann-Whitney U test, which indicates that the lymphocyte density has no significant relationship to the peripancreatic vascular invasion. Fig. 14 presents the distributions of the TLS density and lymphocyte density for the two groups.

IV. Discussion

A. Failure Case Analysis and Limitations

Although the proposed method shows significant superiority in TLSs detection, there are also some failure cases in our experiments. Fig. 15 shows some failure cases where the lymph node areas are mistakenly identified as TLSs because of the high lymphocyte nuclei density. The LDA mechanism can detect the regions with high lymphocyte nuclei density, which is an important feature of the TLSs. However, the lymph node tissue also has a high lymphocyte density feature, which can be one wrong guidance. Further improvement will be included in our subsequent works.

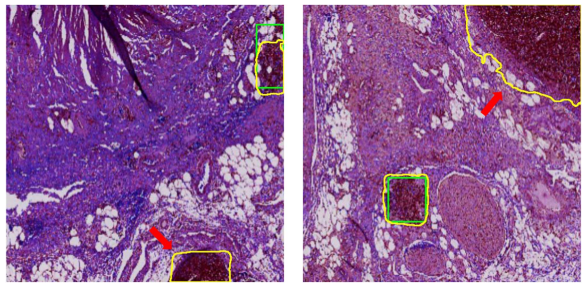

Fig. 15. Failure cases predicted by the proposed method. The segmentation and ground truth are denoted by yellow contours and green rectangles, respectively. The lymph node areas are mistakenly identified as TLSs (signed by red arrows) because of the high lymphocyte nuclei density.

B. Future Directions

The proposed LDA improves the TLS segmentation performance by multi-channel fusion. However, the proposed LDA can also be embedded in the U-shape network architecture or used as an extra loss function for training, which can be studied in our subsequent work.

Regarding the detected TLSs, we have examined the relationship between TLSs and vascular invasion. Through a study of two central patients, we found that TLSs have a stronger association with vascular invasion than individual lymphocytes, highlighting TLSs as an important component of the tumor immune microenvironment. With the expansion of sample size and the subdivision of lymphocyte categories, we will conduct more detailed and in-depth investigations of the immune microenvironment.

As mentioned earlier, the lymphocyte nuclei' morphological characteristics are independent of the organs in which they are located. The lymphocyte nuclei on pathology images of other organs should be morphologically identical to the pancreatic lymphocyte nuclei, although these images have different backgrounds. Therefore, our proposed method can be applied to the TLS detection task of other diseases, such as lung cancer or breast cancer.

V. Conclusion

In this paper, we propose a novel weakly supervised segmentation network embedding cross-scale attention guidance and noise-sensitive constraint for TLS detection. We firstly obtain the segmentation and classification results of the lymphocyte nuclei by combining a pretrained baseline model for nuclei segmentation and a domain adversarial network for lymphocyte nuclei recognition. Then, we establish a cross-scale attention guidance network by jointly learning the coarse-scale features from the original H&E images and fine-scale features from our calculated lymphocyte density attention. A noise-sensitive constraint is introduced by embedding the signed distance function loss in the training procedure to reduce tiny segmentation errors. Experimental results on two collected datasets demonstrate that our proposed algorithm outperforms the state-of-the-art segmentation-based methods for TLS detection. Additionally, we apply our method to validate that the TLS density is significantly related to the peripancreatic vascular invasion based on the clinical data acquired from two independent institutions. Our proposed approach can be applied to the TLS analysis for the tumors in other organs.

References

[1] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, vol. 33, pp. 170–175, 2016.

[2] A. L. Kiemen et al., "Coda: Quantitative 3D reconstruction of large tissues at cellular resolution," *Nature Methods*, vol. 19, pp. 1490–1499, 2022.

[3] J. Zhang et al., "Joint fully convolutional and graph convolutional networks for weakly-supervised segmentation of pathology images," *Med. Image Anal.*, vol. 73, 2021, Art. no. 102183.

[4] G. Xu et al., "Camel: A weakly supervised learning framework for histopathology image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10682–10691.

[5] L. Munoz-Erazo, J. L. Rhodes, V. C. Marion, and R. A. Kemp, "Tertiary lymphoid structures in cancer–considerations for patient prognosis," *Cellular Mol. Immunol.*, vol. 17, no. 6, pp. 570–575, 2020.

[6] C. Sautès-Fridman, F. Petitprez, J. Calderaro, and W. H. Fridman, "Tertiary lymphoid structures in the era of cancer immunotherapy," *Nature Rev. Cancer*, vol. 19, no. 6, pp. 307–325, 2019.

[7] T. N. Schumacher and D. S. Thommen, "Tertiary lymphoid structures in cancer," *Science*, vol. 375, no. 6576, 2022, Art. no. eabf9419.

[8] G. Trajkovski et al., "Tertiary lymphoid structures in colorectal cancers and their prognostic value," *Open Access Macedonian J. Med. Sci.*, vol. 6, no. 10, 2018, Art. no. 1824.

[9] F. Bergomas et al., "Tertiary intratumor lymphoid tissue in colo-rectal cancer," *Cancers*, vol. 4, no. 1, pp. 1–10, 2011.

[10] A. Heindl, I. Sestak, K. Naidoo, J. Cuzick, M. Dowsett, and Y. Yuan, "Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER breast cancer," *JNCI: J. Nat. Cancer Inst.*, vol. 110, no. 2, pp. 166–175, 2018.

[11] K. Mortezaee, "Enriched cancer stem cells, dense stroma, and cold immunity: Interrelated events in pancreatic cancer," *J. Biochem. Mol. Toxicol.*, vol. 35, no. 4, 2021, Art. no. e22708.

[12] A. Redding and E. Grabocka, "A splendid new beginning at the end of a 40-year quest: The first KRASG12D inhibitor in pancreatic cancer," *Cancer Discov.*, vol. 13, no. 2, pp. 260–262, 2023.

[13] S. J. Rubin, R. S. Sojwal, J. Gubatan, and S. Rogalla, "The tumor immune microenvironment in pancreatic ductal adenocarcinoma: Neither hot nor cold," *Cancers*, vol. 14, no. 17, 2022, Art. no. 4236.

[14] S. Graham, Q. D. Vu, E. Shan, A. Azam, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, 2019, Art. no. 101563.

[15] J. Gamper et al., "Pannuke dataset extension, insights and baselines," 2020, *arXiv:2003.10778*.

[16] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[18] C. Shen et al., "Attention-guided pancreatic duct segmentation from abdominal ct volumes," in *Proc. Clin. Image-Based Procedures, Distrib. Collaborative Learn., Artif. Intell. Combating COVID-19 Secur. Privacy-Preserving Mach. Learn.: 10th Workshop, CLIP, 2nd Workshop, DCL 2021, 1st Workshop, LL-COVID19, First Workshop, Tut., PPML, Held Conjunction MICCAI*, 2021, pp. 46–55.

[19] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.

[20] J. Ma, J. He, and X. Yang, "Learning geodesic active contours for embedding object global information in segmentation CNNs," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 93–104, Jan. 2021.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*2015, pp. 234–241.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[23] F. Isensee et al., "nnU-Net: Self-adapting framework for u-net-based medical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[24] E. M. Loyer, C. L. David, R. A. Dubrow, D. B. Evans, and C. Charnsangavej, "Vascular involvement in pancreatic adenocarcinoma: Reassessment by thin-section CT," *Abdom Imag.*, vol. 21, no. 3, pp. 202–206, 1996.

[25] M. Klauss et al., "A new invasion score for determining the resectability of pancreatic carcinomas with contrast-enhanced multidetector computed tomography," *Pancreatology*, vol. 8, no. 2, pp. 204–210, 2008.

[26] K. Teramura, T. Noji, T. Nakamura, T. Asano, and S. Hirano, "Preoperative diagnosis of portal vein invasion in pancreatic head cancer: Appropriate indications for concomitant portal vein resection," *J. Hepato Biliary Pancreatic Sci.*, vol. 23, no. 10, pp. 643–649, 2016.

[27] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality," *Biometrika*, no. 3/4, pp. 591–611, 1965.

[28] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.