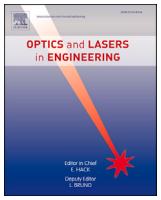
Optics and Lasers in Engineering 184 (2025) 108657

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

journal homepage: www.elsevier.com/locate/optlaseng



Uncertainty-assisted virtual immunohistochemical detection on morphological staining via semi-supervised learning

Shun Zhou [a,b,c,1], Yanbo Jin [a,b,c,1], Jiaji Li [a,b,c], Jie Zhou [a,b,c], Linpeng Lu [a,b,c], Kun Gui [d,e], Yanling Jin [f,g], Yingying Sun [f,g], Wanyuan Chen [f,g,*], Qian Chen [a,c,*], Chao Zuo [a,b,c,*]

[a] *Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China*
[b] *Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210019, China*
[c] *Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, Jiangsu Province 210094, China*
[d] *Konfoong Biotech International CO., LTD., Ningbo, Zhejiang Province 315499, China*
[e] *Ningbo Yangming Medical Inspection Laboratory Co., Ltd., Ningbo, Zhejiang Province 315400, China*
[f] *Cancer Center, Department of Pathology, Zhejiang Provincial Peoples Hospital, Affiliated Peoples Hospital, Hangzhou Medical College, Hangzhou, Zhejiang Province 310014, China*
[g] *Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou, Zhejiang Province 310014, China*

ARTICLE INFO

Keywords:
Bayesian uncertainty
Gastric cancer
p53 immunohistochemistry
Semi-supervision
TP53 gene

ABSTRACT

Tumor suppressor gene TP53 plays a crucial role in cancer diagnosis and prognosis. The gene encodes the tumor suppressor protein p53, which can be identified through immunohistochemical (IHC) staining in various cancers, including gastric carcinoma. However, IHC staining is more costly and therefore not as prevalent as routine hematoxylin-eosin (H&E) staining. In this study, we present a semi-supervised learning-based approach for immunological detection (SSID) of TP53 mutation directly on H&E-stained gastric tissue sections, intending to improve gastric cancer diagnosis. SSID is trained on a small set of annotated image pairs and a larger unannotated dataset of H&E-stained images. It can detect the regions showing strong p53 expression, indicating TP53 mutation, and we validate the accuracy of our approach through both qualitative assessment (pathologists' average score of 2.22/3) and quantitative evaluation (e.g., averaged mean Intersection-over-Union of 0.73). Moreover, we introduce Bayesian uncertainty to assess the credibility of the detected masks, aiming to prevent misdiagnosis and inappropriate treatment. Our results demonstrate that SSID can circumvent the expensive and laborious IHC staining procedures and enable the diagnosis and prognosis of gastric cancer through immunological detection of TP53 mutation.

1. Introduction

With more than one million new cases occurring annually worldwide and dismal overall survival rates, gastric cancer has become one of the most lethal diseases necessitating urgent clinical needs [1]. Histological analysis of stained tissue sections is a critical step in the pathologic evaluation of diseases like gastric cancer. Among the routinely employed staining techniques, hematoxylin-eosin (H&E) staining stands out for its widespread use in clinical pathology and visual inspection of histochemically stained tissue slides [2]. However, it is still challenging to provide a precise pathological diagnosis by using only single-mode staining of H&E since H&E-stained histological slides primarily display fundamental morphological information without reflecting micro-molecular details [3]. Hence, to reveal subcellular features such as cytoplasmic and nuclear details in gastric cancer, researchers have explored various immunohistochemical biomarkers, including nuclear protein Ki67, human epidermal growth factor receptor 2 (HER2) and the tumor suppressor protein p53, all of which enhance cancer analysis [1,4].

Typically, the TP53 gene encodes the p53 protein and has been proven to have high relevance to human cancers, showing great value

* Corresponding authors.
E-mail addresses: chenwanyuan@hmc.edu.cn (W. Chen), chenqian@njust.edu.cn (Q. Chen), zuochao@njust.edu.cn (C. Zuo).
[1] Equal contribution.

https://doi.org/10.1016/j.optlaseng.2024.108657
Received 9 July 2024; Received in revised form 25 September 2024; Accepted 20 October 2024

in cancer diagnosis and prognosis [5]. Moreover, the p53 protein reflects the underlying TP53 mutation status and is closely associated with tumor formation. Numerous therapeutic applications based on TP53 mutation status have been proposed, such as the distinction between two histotypes and pathological predictions within specific histotypes or across multiple histotypes [6]. As one of the gold standards for cancer analysis and diagnostic decisions, immunohistochemical (IHC) staining is an important tool with the capacity to identify specific biomarkers in clinical practice [7]. Combining p53 IHC staining with routine H&E staining significantly improves the diagnostic accuracy of gastric cancer. However, IHC staining requires experienced histotechnologists to perform laborious tissue preparation and chemical processing, which is more costly and time-consuming compared to the standard H&E staining.

Rapid advances in pathologic evaluation of diseases based on deep learning (DL) have emerged in recent years. The deep neural network, for example, the convolutional neural network (CNN) shows a high model capacity to theoretically function arbitrary mappings from one pathological domain to another one [8,9]. Leveraging the assumption of a certain relationship between the latent details of specific biomarkers and the morphological information, deep learning has enabled virtual HER2 IHC staining [7], the prediction of Ki67 positive cells in H&E-stained images [10], and the recognition of epithelial cells for breast cancers [11], bypassing the complicated IHC staining process. Nevertheless, manual annotation is hard to perform, and insufficient training data leads neural networks to learn trivial solutions due to over-fitting. To address these issues, semi-supervised learning (SSL) is adopted due to the ingenious usage of unannotated data to achieve better performance [12]. Furthermore, since neural networks are black-box models, a satisfactory explanation of the network behavior becomes crucial [13], especially when deep learning is applied to clinical application scenarios to avoid misdiagnosis. To render deep learning interpretable, Bayesian inference has recently emerged as a robust method, which offers a mathematically grounded framework to estimate model uncertainty expressing the neural network's confidence in its prediction [13,14]. Bayesian neural networks quantify uncertainties using Monte Carlo dropout or Concrete dropout, replacing the deterministic network weights with probability distributions. The resultant uncertainty maps characterize imperfections that are unknown in real-world applications, such as noise, model error, incomplete training data, and out-of-distribution test data. In the absence of reference data, Bayesian uncertainty provides an effective representation of the error distribution in the inference results. Accordingly, the application of Bayesian models becomes promising, with an ongoing shift in many fields toward utilizing Bayesian uncertainty [15–17].

In this work, we present a novel diagnostic technique called semi-supervised learning-based immunological detection (SSID) for the mutations of the TP53 gene. This method overcomes the limitations of expensive and laborious IHC staining procedures by using deep learning to directly detect p53-positive cells from H&E-stained gastric sections, aiding in gastric cancer diagnosis. It tactfully leverages extra unannotated data to enhance the training of the detection network, albeit with a small quantity of annotated data. To prevent misdiagnosis of TP53 mutation, Bayesian inference was adopted to empower the network and alert pathologists of suspicious regions by estimating the uncertainty. In our experiments, we performed hold-out validations on H&E images, and the accuracy of SSID was validated by both qualitative assessment (pathologists' average score of 2.22/3) and quantitative evaluation (e.g., averaged mean IoU of 0.73). Bayesian uncertainty was then verified to be valid for describing the confidence of SSID in its detection, through the uncertainty analysis of a wrong detection and a correct detection. Beyond its application in p53 protein, SSID has the potential to be extended to other biomarkers and benefit the generalization of unsupervised approaches and uncertainty estimation in immunological applications.

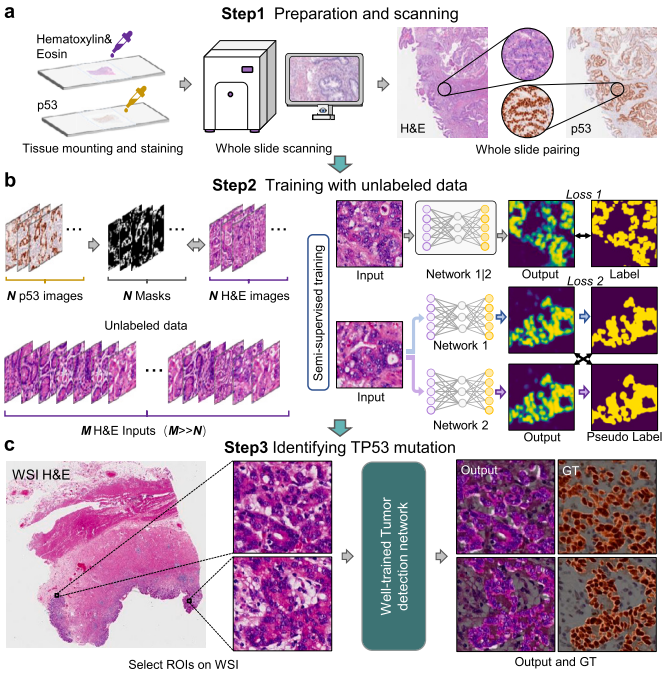


Fig. 1. The integral pipeline of the immunological detection for TP53 mutation based on the semi-supervised method. The schematic outlines the key steps of our proposed technique, including **a**. sample preparation, **b**. network training, and **c**. whole slide image (WSI) detection of TP53 mutation.

2. Principle and methods

2.1. Overview of the detection for TP53 mutation via semi-supervised learning

In the present study, we proposed SSID, a semi-supervised learning-based mutant TP53 detection technique based on H&E-stained input images and masks generated from IHC p53-stained images. The basic pipeline of SSID is illustrated in Fig. 1. First, we performed H&E staining and p53 IHC re-staining on gastric tissues and then mounted them on standard glass slides, followed by slide digitalization and image registration, as shown in Fig. 1a. Next, to improve the low data supply we encounter, we trained the network of SSID according to the semi-supervised approach, as briefly depicted in Fig. 1b, which uses H&E-stained images and the corresponding masks as inputs and ground truths, respectively. Finally, the well-trained network allows us to directly detect TP53 mutations in gastric cancer on H&E-stained images, as shown in Fig. 1c.

2.2. Data processing and construction

Gastric tissue sections were digitally archived as whole slide images (WSIs) by a digital slide scanner. Since H&E and p53 staining were performed on the same section, the H&E-stained WSI and re-stained p53 WSI were naturally in initial alignment as shown in Fig. 2a, implying no obvious morphological discrepancy like serial sectioning. However, the minor offset between H&E and p53 WSI is still present as shown in Fig. 2a at the positions indicated by the black arrows, and thus further registration was performed on these coarsely aligned image pairs. The first step was to crop an image patch with a resolution of 512×512 pixels (0.12 mm×0.12 mm) from the H&E-stained WSI as the reference and search the roughly corresponding area (600×600 pixels) of the p53-stained WSI for a matched patch.

Subsequently, as shown in Fig. 2b, a score matrix of structural similarity index measure (SSIM) [18] was produced by correlating each H&E-stained patch with the p53-stained image searched in pixel-by-pixel steps within the corresponding area of the p53 WSI, and the entry

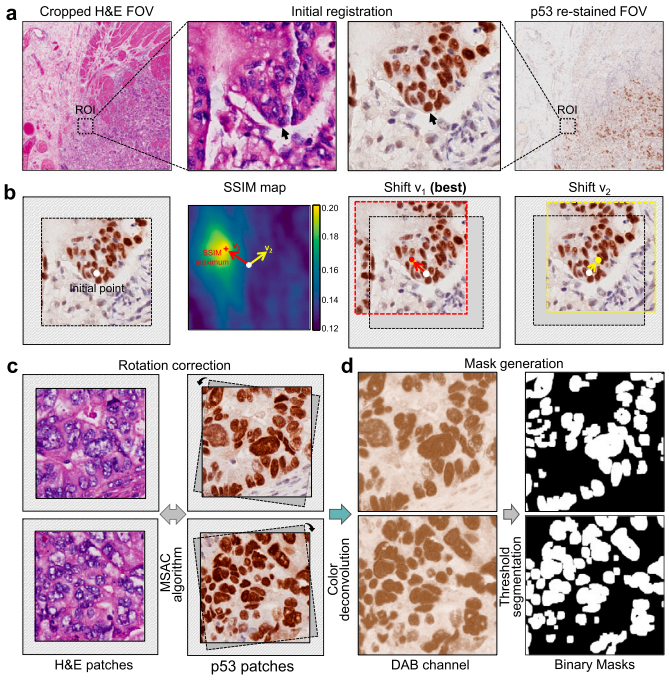


Fig. 2. The pipeline of the image registration and mask generation. **a**. Image pairs of H&E and p53 staining get initially registered after re-staining the same gastric tissue. **b**. Further alignment is then achieved by determining the location of the maximum value of the SSIM score matrix. **c**. Rotation correction is performed for more precise registration by using the SURF and MSAC algorithms. **d**. Mask generation is implemented by progressively applying DAB extraction, threshold segmentation, erosion, and dilation on p53-stained images.

with the maximum SSIM score can represent the most matched patch in the p53-stained WSI. The SSIM index between two images is calculated as:

$$SSIM(I_1, I_2) = \frac{(2\mu_1\mu_2 + c_1)(2\sigma_{1,2} + c_2)}{(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2 + \sigma_2^2 + c_2)}, \tag{1}$$

where μ_1 and μ_2 are the means of the two images I_1, I_2. σ_1 and σ_2 are the standard deviations, $\sigma_{1,2}$ is the mutual covariance, and c_1 and c_2 are the regularization parameters. Thereafter, to correct the possible rotation angle between H&E and p53 images, we applied the transformation matrix calculated by using the speeded-up robust features (SURF) algorithm [19] and the M-estimator sample consensus (MSAC) algorithm [20] to p53-stained images, as shown in Fig. 2c. To efficiently generate a ground-truth mask indicating TP53 mutation and reduce manual annotation costs, we performed color deconvolution on p53-stained images to extract the diaminobenzidine (DAB) channel [21]. This process allowed us to obtain masks covering the regions of mutant TP53 by progressively applying threshold segmentation, erosion, and dilation on the DAB channel images of p53 (Fig. 2d).

2.3. Semi-supervised network training and uncertainty estimation

In this section, we introduce the network architecture used in SSID. Unlike the CNN with convolution kernels that only focus on the local receptive field, Transformer leverages a self-attention mechanism for efficient modeling of both long and short-range spatial interactions [22]. For the feature extraction of underlying immunological modality in the H&E image, a larger receptive field means access to a larger field of view and more complete morphological information of the H&E-stained tissue structure. Therefore, as shown in Fig. 3a, the network of SSID utilizes a Transformer-based encoder to capture global or long-range dependencies. Supposing that the network consists of L layers, $\theta = \{\mathbf{\Theta}_l\}_{l=1}^{L}$ signifies the deterministic weights and $\omega = \{\mathbf{W}_l\}_{l=1}^{L}$ represent the set of random weight matrices after applying Concrete dropout [14] on θ to form a Bayesian neural network. It is worth noting that the selected Concrete dropout method enhances the efficiency of Bayesian deep learning by allowing automatic adjustment of the dropout probabilities along with the optimization of model parameters. Compared to other uncertainty estimation methods, such as Monte Carlo dropout and Deep ensembles, this method offers higher computational efficiency. A random weight matrix $\mathbf{W}_l$ (the shape is $C_l \times C_{l-1}$), decorated by Concrete dropout module, can be expressed as:

$$\mathbf{W}_l = \mathbf{\Theta}_l \cdot \mathrm{diag}\left([z_l^{(1)}, z_l^{(2)}, ..., z_l^{(C_{l-1})}]\right), \tag{2}$$

where z_l is sampled from Bernoulli distribution, i.e., $z_l \sim B(1, p_l)$. The encoder first performs patch partition on the H&E input image to form non-overlapping patches and then carries out feature extraction through four stages of Swin Transformer blocks [23] (Fig. 3b). The Swin Transformer blocks are cascaded residual structures characterized by a window-based multi-head self-attention module (W-MSA), a shifted W-MSA (SW-MSA), layer normalization (LN) and multi-layer perceptrons (MLP). A linear embedding layer is applied before the first stage and patch merging layers are arranged before subsequent stages. Besides the encoded features at the end of Swin Transformer, the shallow features extracted in the first stage skip to the decoding branch. The decoder (Fig. 3a) adopts the structure of deeplabv3+ [24] which has been widely applied and proven effective. It is responsible for feeding the encoded latent features into its Atrous Spatial Pyramid Pooling (ASPP) module that uses dilated convolutions with different dilation rates to capture multi-scale contextual information. The concatenation of the post-ASPP features and the shallow features from the encoding branch is finally mapped to a binary mask that corresponds to the region with TP53 mutations. Notably, as shown in Fig. 3b, extra Concrete dropout (CD) layers are assembled at the appropriate positions in the network for the estimation of Bayesian uncertainty.

As shown in Fig. 3c, the overall optimization objective for the network training is composed of three terms:

$$\mathcal{L} = \mathcal{L}_s + \lambda_1 \mathcal{L}_u + \lambda_2 \mathcal{L}_{cd}, \tag{3}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ represent the supervision loss on annotated data and the unsupervised loss on unannotated data, respectively. $\mathcal{L}_{cd}$ is the regularization term for Bayesian uncertainty. λ_1 and λ_2 are the weight factors of $\mathcal{L}_u$ and $\mathcal{L}_{cd}$, respectively. To implement semi-supervised learning, we adopted the concept of cross-pseudo supervision [25]. Our semi-supervised framework employs two parallel networks that share the same architecture, as previously described. During the training phase, the optimization algorithm independently updates the weights θ_1 and θ_2 of the two models. We assume that $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$ is the annotated set of H&E input images and corresponding TP53 mutation masks, where $\mathbf{x}_i$ and $\mathbf{m}_i$ are the i-th input and label, respectively. N is the size of the set. $\mathcal{U} = \{\mathbf{u}_i\}_{i=1}^{M}$ is a much larger unannotated set of H&E-stained images, where $\mathbf{u}_i$ denotes the i-th H&E image without annotation. Based on the symbolic definition above, we express the cross entropy $\mathcal{L}_{ce}$ between two probability distributions $\mathbf{t}$ and $\mathbf{s}$ as:

$$\mathcal{L}_{ce}(\mathbf{t}, \mathbf{s}) = -\sum_{c=0}^{C-1} t(c) \log(s(c)), \tag{4}$$

where C stands for the number of categories, which is 2 for SSID. $t(c)$ and $s(c)$ are the values of $\mathbf{t}$ and $\mathbf{s}$ at c-th category, respectively.

At each training iteration, each batch includes B annotated H&E images from $\mathcal{X}$ and B unannotated ones from $\mathcal{U}$. Supposing that the two networks establish two mappings, f_1 and f_2, the supervision loss function $\mathcal{L}_s$ can be expressed as:

$$\mathcal{L}_s = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{ce}(\mathbf{m}_i, \sigma(f_1(\mathbf{x}_i))) + \mathcal{L}_{ce}(\mathbf{m}_i, \sigma(f_2(\mathbf{x}_i))), \tag{5}$$

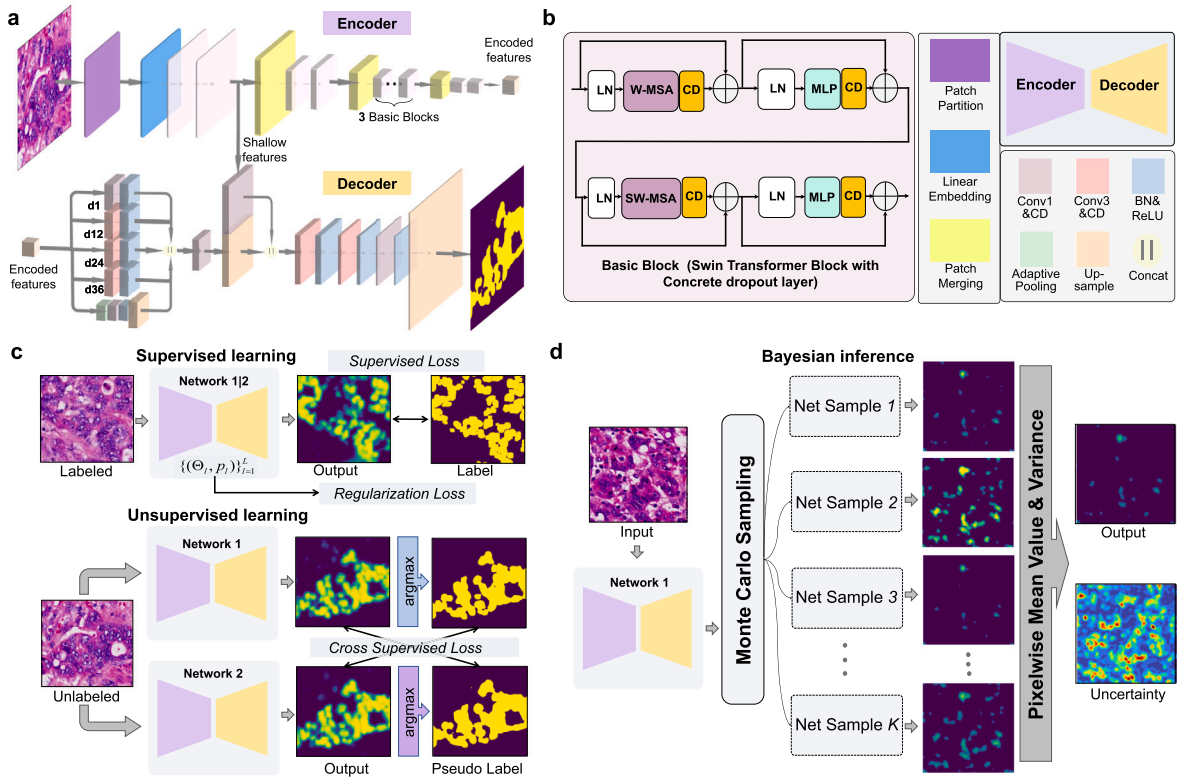


Fig. 3. The illustration of network training via semi-supervised learning. The SSID's network includes an encoder and a decoder, endowed with Concrete dropout (CD) layers. **a**. The encoder mainly consists of four stages of Swin Transformer blocks, and the shallow features extracted in the first stage skip to the decoding branch, a deeplabv3+ network with an ASPP module to capture multi-scale contextual information. **b**. The structure of Swin Transformer block with Concrete dropout layer. **c**. The semi-supervised learning scheme. **d**. The schematic diagram of uncertainty prediction via Bayesian inference.

where $\sigma(\cdot)$ denotes the softmax activation function, i.e., $\sigma(\mathbf{s}) = e^{s(c_t)} / \sum_{c_s}^{C-1} e^{s(c_s)}$.

For each of the unannotated H&E images in the sampled batch, we pass it through two networks simultaneously. Its predicted class distribution map $\mathbf{p}_{ki} = f_k(\mathbf{u}_i)$ is then treated as pseudo-label by using the formula $\hat{\mathbf{m}}_{ki} = \arg\max_c p_{ki}(c)$. By imposing cross-supervision with pseudo labels, one network generates pseudo labels to supervise the other using the standard cross-entropy loss, and vice versa. This process effectively performs network perturbation, which implicitly expands the data distribution of the training set [25]. The cross-supervision loss function on the unannotated data can be written as:

$$\mathcal{L}_u = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{ce}\left(\hat{\mathbf{m}}_{2i}, \sigma\left(f_1\left(\mathbf{u}_i\right)\right)\right) + \mathcal{L}_{ce}\left(\hat{\mathbf{m}}_{1i}, \sigma\left(f_2\left(\mathbf{u}_i\right)\right)\right). \tag{6}$$

We further confer on SSID the ability to infer the model uncertainty of detected results. Specifically speaking, we adopted Concrete dropout modules to approximate the posterior distributions $q_\theta(\omega)$ of parameters in the network [14]. The loss function for the Concrete dropout-equipped network can be defined as regularization for network parameters and dropout rate:

$$\begin{aligned}\mathcal{L}_{cd} = \mathbb{KL}\left(q_\theta(\omega) \| p(\omega)\right) &= \sum_{l=1}^{L} \mathbb{KL}\left(q_{\Theta_l}\left(\mathbf{W}_l\right) \| p\left(\mathbf{W}_l\right)\right) \\ &\propto \sum_{l=1}^{L}\left(\frac{l^2\left(1-p_l\right)}{2}\|\Theta_l\|^2 - \gamma\mathcal{H}\left(p_l\right)\right),\end{aligned} \tag{7}$$

where the Kullback-Leibler ($\mathbb{KL}$) divergence is used to measure the discrepancy between the posterior distribution $q_\theta(\omega)$ and the prior distribution of the weights $p(\omega)$ that is assumed as Gaussian distribution here. p_l is a tunable dropout probability of each layer and γ is a constant to balance the two regularization terms in Eq. (7): the first term functions as the regularization of the weight matrices, and $\mathcal{H}(p_l)$ in the second term is $-p_l log p_l - (1-p_l) log(1-p_l)$ the entropy of a Bernoulli random variable with probability p_l. To render Bayesian neural network conveniently trainable, Concrete dropout module introduces Concrete distribution relaxation to reparameterize discrete Bernoulli distribution, thus the sampled variable can be formulated as:

$$z_l = \text{sigmoid}\left(\frac{1}{t}\left(\log\frac{p_l}{1-p_l} + \log\frac{u}{1-u}\right)\right), \tag{8}$$

where u satisfies uniform distribution, i.e., $u \sim \text{Unif}(0,1)$, and t denotes a temperature variable that restricts the values to the interval from 0 to 1. Once obtaining a well-trained Concrete dropout-based network, we can employ the Monte Carlo (MC) integration over K samples to carry out random dropout sampling of weight matrices, and the predictive variance [13,26] is used for Bayesian inference (Fig. 3d) to measure the model uncertainty of TP53 mutation detection:

$$\begin{aligned}\text{Var}\left(\mathbf{p}^* \mid \mathbf{x}^*, \mathcal{X}\right) \approx &\frac{1}{K}\sum_{k=1}^{K}\sigma\left(f^w\left(\mathbf{x}^*\right)\right)^\top \sigma\left(f^w\left(\mathbf{x}^*\right)\right) \\ &-\frac{1}{K^2}\sum_{k=1}^{K}\sigma\left(f^w\left(\mathbf{x}^*\right)\right)^\top \sum_{k=1}^{K}\sigma\left(f^w\left(\mathbf{x}^*\right)\right),\end{aligned} \tag{9}$$

where $\mathbf{x}^*$ represents an unseen H&E-stained image and $\mathbf{p}^*$ is the predicted probabilities of TP53 mutation.

2.4. Materials and experiment implementation details

The training materials for this study were selected from the most invasive sections of 65 tissue samples of high-grade gastric adenocarcinoma. The samples were obtained from the Laboratory of Endocrine Gland Diseases and were prepared by the Ningbo Yangming Medical Inspection Laboratory. The gastric tissue sections of patients were ob-

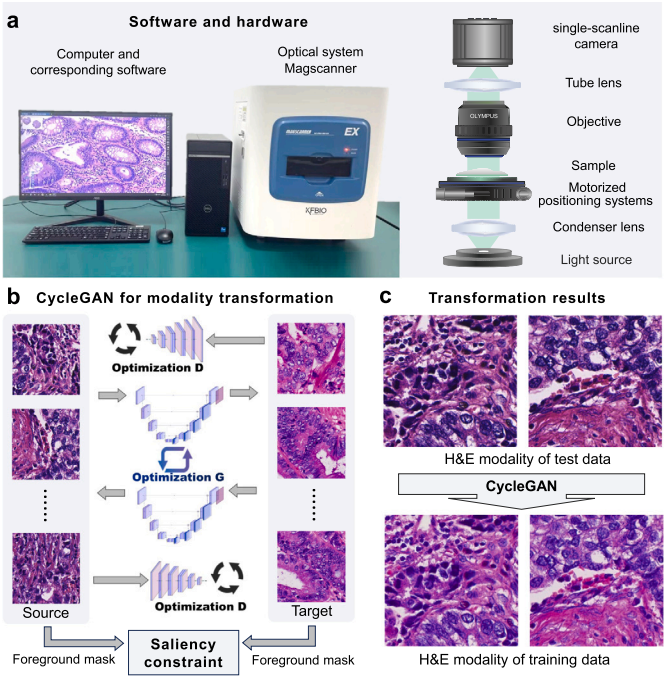


Fig. 4. a. The experimental instrument of digital pathology scanner and the schematic illustration of internal imaging system. **b**. The CycleGAN scheme for domain transformation of H&E modalities. **c**. The CycleGAN converts the distribution of the test data to the same as the training data.

tained under Ethics Committee Approval (Zhejiang Provincial People's Hospital Ethics Review 2022 Approval No.074). All the tissue samples were guaranteed to be de-identified of patient-related information under supervision. In addition, the failure of standard IHC staining or severe tissue damage is a norm even in experienced pathology laboratories [7], and the related waste also occurred in our specimen preparation with about one-third of the tissue sections discarded eventually. It is worth noting that we adopted two main patterns of p53 staining, overexpression and cytoplasmic as ground truth, and the two patterns suffice for the strong prediction of mutation-type p53 in the vast majority of cases [6]. In the general staining procedure, multi-stained sections of a certain region are obtained by staining serial sections. However, sequential sections may skim over the critical areas targeted for examination, resulting in the loss of areas needed for critical diagnosis, and hence we performed p53 immunostaining on the same tissue sections via the re-staining method [27]. The specific steps for re-staining are described below. First, we performed formalin-fixing paraffin-embedding (FFPE) and H&E staining procedures [2] on tissue sections. Next, the as-prepared H&E slides were manually rinsed (30-40 dips in reagent) in slide baskets and soaked in Acetone for the effective removal of the coverslip. Then the slides were put in the xylene bath with 3 times rinses and subsequently in 95% alcohol with 3 minutes hold intervals for approximately 30 minutes to remove the eosin stain, during which the covering of paraffin film should be guaranteed to reduce alcohol evaporation and expedite eosin removal. After the washing of slides in the deionized water, the reaction buffer was employed to remove hematoxylin from slides with 3-5 rinses. After waiting for the slides to dry in the hood for 5 minutes, we finally conducted antibody IHC assay detection protocols with antibodies against p53. Noticeably, all procedures should be operated in a fume hood by using manual wash stations, negating the need for heat that may damage the tissue.

We scanned the tissue sections under the KFBIO digital pathology scanner KF-PRO-005-EX, and its appearance (including software and hardware) and internal optical imaging system structure are present in Fig. 4a. The acquired WSIs were cropped into small patches (512×512 pixels, 0.12 mm×0.12 mm) to accommodate the memory of the graphic processing unit (GPU). Then we manually screened the acquired dataset to exclude training materials with obvious defects (e.g., staining failure, and defocused images). The dataset of gastric tissues was further divided into a training set and a test set. Since mutation-type p53 patterns are rarer than wild-type, the dataset turns smaller after further selection of data containing mutant TP53. The final training set contains 213 annotated pairs and 3895 unannotated H&E images, whereas the test set consists of 768 H&E images independent from the training data. Note that the unannotated H&E images are from a gastric tissue section without corresponding registered p53-stained label, and as shown in Fig. 4b, they are transformed to approximate the H&E modality of training data (Fig. 4c) by using a cycle-consistent generative adversarial network (CycleGAN) with saliency constraint [28]. This preprocess ensures that the distribution of test data is consistent with that of the training data and avoids the variation among staining approaches [29]. This study was conducted on a workstation equipped with an Intel i9-7980XE 18-core 2.60 GHz CPU (128 GB RAM) and an NVIDIA GeForce RTX 3090 GPU, by using Python 3.7 and PyTorch 1.8.0. The adopted optimization algorithm is stochastic gradient descent (SGD) [30], with a learning rate of 0.0008, the momentum set to 0.9, and the rest left at default values. The training of the SSID network required approximately 10 hours for 100 epochs. For individual-sample inference, the outputs of SSID include a predicted mask of TP53 mutations and a corresponding data uncertainty map. The average processing time was 1.16 seconds for small patches (512×512 pixels), and 56.83 seconds for large patches (2024×2048 pixels).

3. Experimental results

3.1. Whole-slide-level and patch-level presentations for TP53 mutation detection

We demonstrated SSID's inference by feeding it with H&E-stained images that never appeared in the training stage. Fig. 5a summarizes a WSI-level example of an H&E-stained gastric gland tissue that was diagnosed as high-grade carcinoma. As shown in Fig. 5b, a whole field detection for H&E-stained tissue within a dashed box was inferred, which covers the regions with a high probability of TP53 mutation (white area) and corresponds to the strong diffuse nuclear p53 over-expression. Furthermore, we selected three regions of interest (ROIs) that contain tumor cells (areas stained dark brown by DAB chromogen) and one region that does not contain tumor cells, and enlarged them in Fig. 5c for a more detailed presentation. The first row of Fig. 5c shows the H&E-stained images of the four ROIs, which are inputs of SSID's network, and the second row shows their p53-stained counterpart. The third row shows the SSID-detected masks overlaid with their H&E-stained counterparts in the first row, and the fourth row shows the ground-truth (GT) masks overlaid with their IHC p53-stained counterparts in the second row. Apparently, the detections of mutant TP53 shown in the last two rows are consistent with each other, faithfully in accord with the abnormal p53 expression patterns shown in the second row. It is intuitively clear that SSID is capable of detecting lesioned regions with mutant TP53. Furthermore, SSID can also provide accurate inferences for non-lesion areas, thereby avoiding misdiagnosis.

3.2. Uncertainty estimation of the detected TP53 mutation

To evaluate the efficacy of Bayesian uncertainty in revealing SSID's error in the absence of ground truth, we selected two examples for Bayesian inference of SSID (Figs. 6a and 6c). Figs. 6a and 6b illustrate a case of false detection, while Figs. 6c and 6d illustrate a case of true detection. Notably, this false detection case is one of the rare, significant detection errors among the 768 results. In both examples, a patch-level H&E-stained image serves as input to SSID's network. The network detects the region with TP53 mutations, matching the ground-truth region identified by the p53 staining pattern. Figs. 6a and 6c display the actual

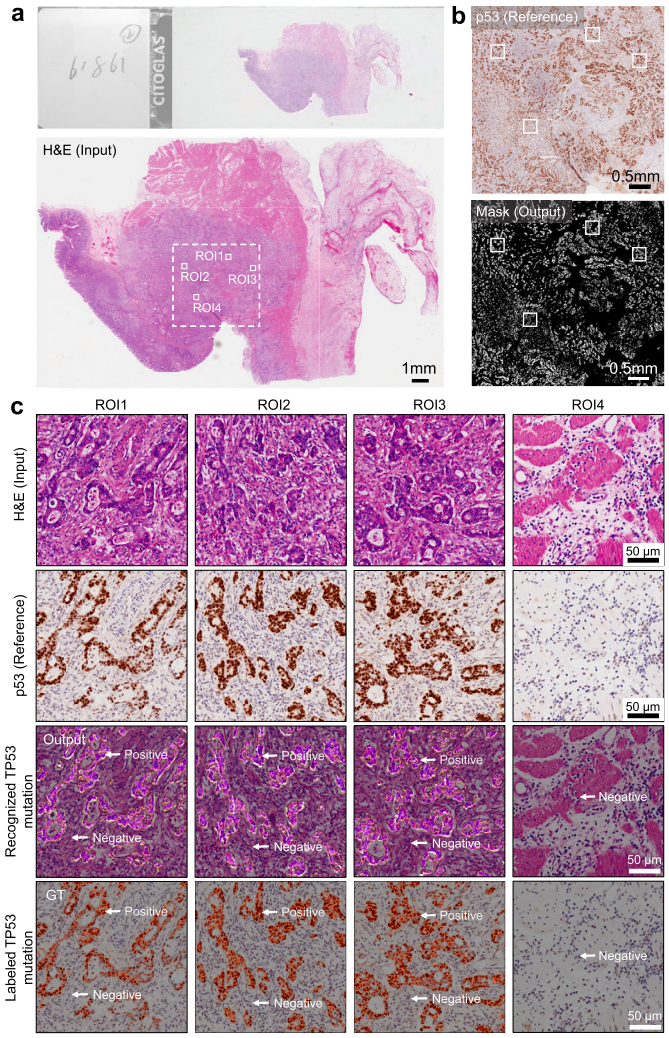


Fig. 5. The illustration of TP53 mutation detection at WSI level and patch level. **a**. An example of H&E-stained WSI. **b**. The corresponding p53 staining pattern within the dashed box in (**a**) and the TP53 mutation mask detected by SSID. **c**. The results from the first to fourth columns correspond to ROI1-ROI4 outlined in the dashed box of (**a**). For each ROI, images from the first row to the fourth row represent the H&E-stained input image, the IHC p53-stained counterpart, the detected mask of mutant TP53, and the GT mask generated based on the p53-stained image, respectively.

TP53 mutation and the recognized mutation, along with their corresponding masks. The error map and the predicted uncertainty of both examples are shown in Figs. 6b and 6d. Note that the error map is defined as the pixel-level discrepancy between the recognized mask and the actual mask. For detailed comparisons between the error map and the predicted uncertainty, we selected two ROIs and presented their enlarged views in Figs. 6b and 6d. In this context, uncertainty indicates that the network's output may not be entirely reliable and may require manual re-evaluation. Higher uncertainty values suggest that the network is unsure whether a region is healthy or diseased. In Fig. 6b, which pertains to the false detection example, ROIs A and B show that the undetected cells are assigned high uncertainty values, approximately ranging from 0.5 to 0.8, matching well with the distribution of large error. We preliminarily suspect that this detection failure is caused by certain specific characteristics inherent to current pathological slides. Due to the rapid proliferation of cancer cells, they tend to densely aggregate, forming stacked or mosaic patterns. As a result, sparsely distributed cancer cells (Fig. 6a) may not be detected due to their absence or limited presence in the training set during network training, leading to misidentification. In contrast, Fig. 6d, which pertains to the true detection example, shows that ROIs C and D exhibit low uncertainty values, mostly ranging from 0.1 to 0.4. The uncertainty maps in these regions are sparsely distributed and primarily located at the cell profiles, accurately reflecting the error map.

3.3. The comparison between semi-supervision and full supervision

Since the network of SSID was trained through the semi-supervised learning strategy, it is necessary to show the superiority of the semi-supervised method over the fully supervised method. We compared their performances using the mean Intersection-over-Union (mIoU) metric averaged on the test data. Unlike the supervised baseline, the semi-supervised learning method additionally exploits 3895 unannotated H&E images to implicitly expand the training data. mIoU is a standard evaluation metric used in image segmentation tasks, which measures the degree of coincidence between the segmentation and the ground truth, defined as the ratio of the intersection to the union of the detected and ground-truth regions for each class. After training networks for the same iterations through semi-supervised learning and supervised learning, we tested the performances of the two methods on the test data. Fig. 7a shows a line graph in which the semi-supervised method consistently outperforms the supervised baseline almost at each epoch whether using Swin Transformer [23] or ResNet-101 [31] as the backbone network. The maximum mIoU values of the four methods across all epochs are illustrated in Fig. 7b, indicating that Swin Transformer under semi-supervision performs best among the four combinations.

3.4. Performance analysis

We tested SSID's performance from both qualitative and quantitative perspectives. Considering our intention to provide a valid reference for clinical diagnosis, two board-certified pathologists were invited to blindly evaluate the TP53 mutation masks detected by SSID's network. In line with the matching degree between the detected mask and the ground truth, an assessment criterion was given on a scale of 0 to 3, with 0 representing complete failure, 1 representing only partial agreement, 2 representing an acceptable agreement, and 3 representing high agreement with negligible differences. Two pathologists scored the 768 detection results and counted the number of results for each score. Fig. 7c summarizes human scoring of TP53 mutation masks, with an average score of 2.10 (σ = 0.88) set by the first pathologist and 2.34 (σ = 0.78) set by the second pathologist. Moreover, the IoU values of the 768 results were categorized into four levels based on upper quartile, median, and lower quartile, corresponding to the four ranks of the pathologists' scores. The distribution of test results in each level was 116, 263, 257, 132. We conducted chi-square test between the IoU ranking and the pathologists' rating scores, yielding chi-square values of 128.52 and 136.75, both exceeding 16.92 (the degree of freedom of 9). This indicates a p-value of less than 0.05, suggesting that there is no statistically significant difference and the results of SSID are consistent with the pathologists' assessments.

Next, the quantitative evaluation of the mutant TP53 masks was carried out by using a series of widely-used metrics in medical analysis. Besides the already-employed Intersection-over-Union (IoU) in the previous section, we also invoked other metrics such as F1-score, Acc (Accuracy), Spec (Specificity), and Sens (Sensitivity), which are all quantitative metrics used to evaluate the performance of a binary classification model. The detected masks were compared to the p53-based ground truths by assessing these metrics (mean F1 = 0.7020, mean Acc = 0.8884, mean Spec = 0.9277, mean Sens = 0.7195, mean IoU = 0.7327), which were averaged over all images of the test set. Fig. 7d presents the boxplot of these metrics and Fig. 7e shows the ROC curve (the average AUC is 0.83) with FPR (False Positive Rate) as the horizontal coordinate and TPR (True Positive Rate) as the vertical coordinate.

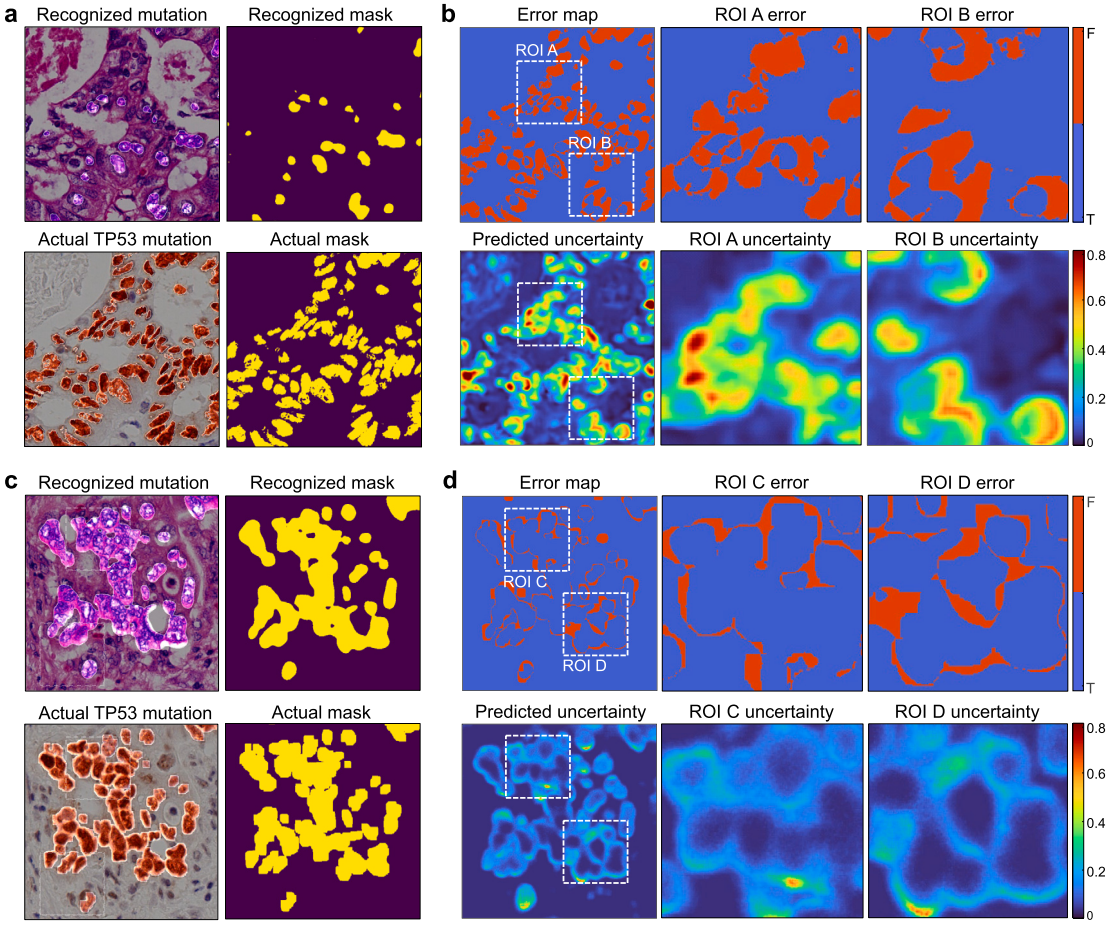


Fig. 6. Uncertainty analysis with over-expression pattern of p53 staining for two different examples, illustrating both correct and false detection of TP53 mutations. **a**, **b**. Display the H&E-stained image with overlaid recognized TP53 mutation regions alongside their corresponding ground truth generated based on the p53-stained images (first column in (**a**)), and their associated masks (second column in (**a**)). Additionally, (**b**) presents the error map (first row) and predicted uncertainty (second row) for the false detection case. **c**, **d**. Show the corresponding results for the correct detection of TP53 mutations.

The ROC curve is a graphical representation of the performance of a binary classifier system and the AUC curve is a single scalar value that summarizes the performance of the ROC curve. Moreover, the true positive, false positive, and false negative of the detected mask were overlapped with the ground-truth mask in color-coded form as illustrated in Figs. 8a and 8b to provide a more detailed visualization of TP53 SSID's performance.

4. Conclusions and discussion

In conclusion, we have developed SSID, a learning-based biomedical identification technique aimed at the immunological detection of TP53 mutations by incorporating semi-supervision and Bayesian inference for the diagnosis of gastric cancer. The proposed SSID framework takes advantage of the unannotated dataset of H&E images to detect TP53 mutations and achieves enhanced performance despite the unavailability of massive and valid IHC data for ordinary cases. The noise introduced by network perturbation in semi-supervision can implicitly expand the training set, alleviating the common issue where new samples deviate significantly from the distribution modeled by the limited data, leading to incorrect predictions. By using the MC Concrete dropout approach, SSID can infer pixel-wise uncertainty maps to quantify the reliability of the network on its detection and explain the prediction error caused by the small data volume. Specifically, Bayesian uncertainty helps identify incorrect detection of TP53 mutations by assigning high uncertainty, as well as make correct detection convincing by assigning low uncertainty.

Notably, our proposed method for immunological diagnosis offers greater convenience compared to previous modality transformation techniques, such as virtual staining [7,32,33] and stain transformation [34]. This advantage is primarily due to the fact that visual assessments of immunological patterns in both standard and digital staining are user-dependent and prone to diagnostic errors caused by staining imperfections. Besides, the binary masks for TP53 mutation detection contain less redundant information than staining modalities, making it easier to train an accurate network.

Despite the advantages and benefits brought by our proposed technique, we cannot ignore remaining challenges of enhancing the diversity and quality of data. Considering artificial maloperations and interlaboratory variations during slide production, staining, and sample scanning procedures, the unsatisfactory conditions of H&E and p53 images is inevitable, which necessitates an automated data cleansing solution to eliminate inferior training data and avoid misleading clinical inputs. Additionally, it should be noted that our dataset only covers two major mutation-type p53 patterns excluding another mutation-type case of complete absence caused by splice site mutations or truncating mutations [6]. Although this special case shows a normal wild-type staining pattern, it is actually a mutation type in the clinical diagnosis, resulting in the failure of the automatic generation process of ground-truth masks for this mutation-type pattern. Thus, the distribution of our dataset is not sufficiently diverse and may deviate from the real distribution to a certain extent, probably affecting the generalizability of the detection network.

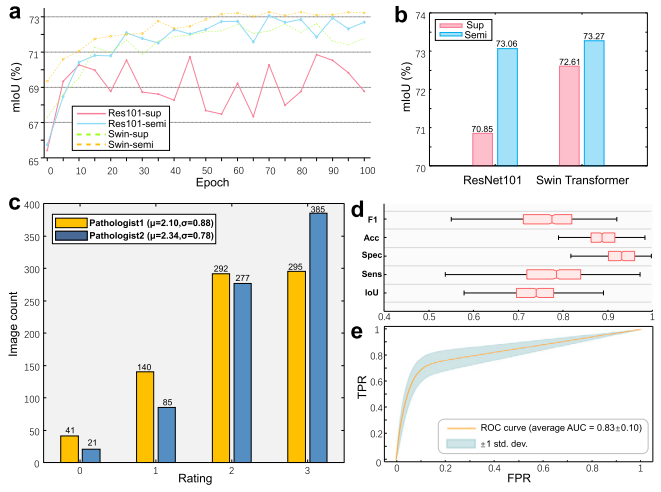


Fig. 7. The comparison between semi-supervision (SSID) and full supervision, the visual evaluation based on the predetermined criterion, and the metrics-based assessment of TP53 mutation regions detected by SSID. **a**. The semi-supervised network (Semi) was compared with the supervised baseline (Sup) by using ResNet-101 (Res101) and SwinTransformer (Swin) as the backbone, respectively, over 100 training epochs. **b**. Performance comparison with different learning schemes and network architectures. **c**. Pathologist scoring for 768 detection results. **d**. Six metrics for medical evaluation, including IoU, Acc, etc., were employed to objectively determine the difference between the detected mask and the ground truth, which is displayed as a boxplot. **e**. The average Receiver Operating Characteristic (ROC) curve with high Area Under Curve (AUC) value reflects the detection capability.

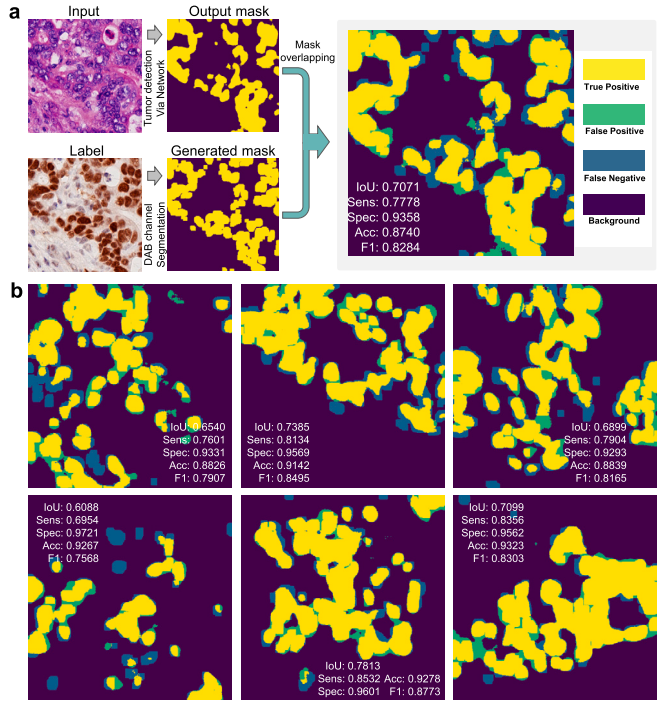


Fig. 8. Detailed visualization of the discrepancy between the detected mask and the ground truth. **a**. The H&E-stained image is input to SSID's network, and the detected mask of TP53 mutation is overlapped with the ground truth to clearly visualize their difference. **b**. Six examples are additionally demonstrated with quantitative assessment results attached.

In the future, we will need to promote the diversification of staining patterns and acquire a large and diverse set of training data to reliably ensure the high stability and generalization ability of the detection technique. To meet the growing demand for the accurate and rapid diagnosis of various cancers at a lower cost, it is necessary to develop a more generalized immunological detection technology for other pathological modalities or biomarkers. Since other less costly imaging methods can also provide morphological information, such as label-free phase imaging [35–37] or diffraction tomography [38–42], we may consider them as alternatives to the H&E staining technique. On top of semi-supervised learning, zero-shot learning [43] can be further introduced to expand the applicability of SSID across the entire range of tumors.

CRediT authorship contribution statement

Shun Zhou: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization, Formal analysis. **Yanbo Jin:** Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization, Formal analysis. **Jiaji Li:** Writing – original draft. **Jie Zhou:** Writing – original draft. **Linpeng Lu:** Writing – original draft. **Kun Gui:** Validation, Methodology, Data curation, Writing – original draft. **Yanling Jin:** Methodology, Formal analysis, Data curation, Writing – original draft. **Yingying Sun:** Data curation, Formal analysis, Writing – original draft, Methodology. **Wanyuan Chen:** Data curation, Writing – original draft, Formal analysis, Methodology, Validation. **Qian Chen:** Funding acquisition, Supervision, Writing – original draft. **Chao Zuo:** Funding acquisition, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFA1205002, 2024YFE0101300), National Natural Science Foundation of China (62361136588, 62105151, 62175109, U21B2033, 62227818), Leading Technology of Jiangsu Basic Research Plan (BK20192003), Youth Foundation of Jiangsu Province (BK20210338), Biomedical Competition Foundation of Jiangsu Province (BE2022847), Key National Industrial Technology Cooperation Foundation of Jiangsu Province (BZ2022039), Fundamental Research Funds for the Central Universities (30920032101, 30923010206), Fundamental Research Funds for the Central Universities (2023102001), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105, JSGP202201).

Data availability

Data will be made available on request.

References

[1] Sexton RE, Al Hallak MN, Diab M, Azmi AS. Gastric cancer: a comprehensive review of current and future treatment strategies. Cancer Metastasis Rev 2020;39:1179–203.
[2] Fischer AH, Jacobson KA, Rose J, Zeller R. Hematoxylin and eosin staining of tissue and cell sections. Cold Spring Harb Protoc 2008;2008. pdb-prot4986.
[3] Fox H. Is h&e morphology coming to an end? J Clin Pathol 2000;53:38–40.
[4] Al-Moundhri M, et al. The prognostic significance of p53, p27kip1, p21waf1, her-2/neu, and ki67 proteins expression in gastric cancer: a clinicopathological and immunohistochemical study of 121 arab patients. J Surg Oncol 2005;91:243–52.
[5] Petitjean A, Achatz M, Borresen-Dale A, Hainaut P, Olivier M. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene 2007;26:2157–65.
[6] Köbel M, Ronnett BM, Singh N, Soslow RA, Gilks CB, McCluggage WG. Interpretation of p53 immunohistochemistry in endometrial carcinomas: toward increased reproducibility. Int J Gynecol Pathol 2019;38:S123.
[7] Bai B, et al. Label-free virtual her2 immunohistochemical staining of breast tissue using deep learning. BME Front 2022;2022.
[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[9] Bakator M, Radosav D. Deep learning and medical diagnosis: a review of literature. Multimodal Technol Interact 2018;2:47.
[10] Liu Y, et al. Predict ki-67 positive cells in h&e-stained images using deep learning independently from ihc-stained images. Front Mol Biosci 2020;7:183.
[11] Valkonen M, et al. Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for er, pr, and ki-67. IEEE Trans Med Imaging 2019;39:534–42.
[12] Yang X, Song Z, King I, Xu Z. A survey on deep semi-supervised learning. IEEE Trans Knowl Data Eng 2022.
[13] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proc Int Conf Mach Learn. PMLR; 2016. p. 1050–9.
[14] Gal Y, Hron J, Kendall A. Concrete dropout. Adv Neural Inf Process Syst 2017;30.
[15] Xue Y, Cheng S, Li Y, Tian L. Reliable deep-learning-based phase imaging with uncertainty quantification. Optica 2019;6:618–29.
[16] Upadhyay U, Sudarshan VP, Awate SP. Uncertainty-aware gan with adaptive loss for robust mri image enhancement. In: Proc IEEE/CVF Int Conf Comput Vis (ICCV); 2021. p. 3255–64.
[17] Mukhoti J, Gal Y. Evaluating Bayesian deep learning methods for semantic segmentation. Preprint. arXiv:1811.12709, 2018.
[18] Sara U, Akter M, Uddin MS. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. J Comput Commun 2019;7:8–18.
[19] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (surf). Comput Vis Image Underst 2008;110:346–59.
[20] Torr PH, Zisserman A. Mlesac: a new robust estimator with application to estimating image geometry. Comput Vis Image Underst 2000;78:138–56.
[21] Shu J, Dolman G, Duan J, Qiu G, Ilyas M. Statistical colour models: an automated digital image analysis method for quantification of histological biomarkers. Biomed Eng Online 2016;15:1–16.
[22] Naseer MM, Ranasinghe K, Khan SH, Hayat M, Shahbaz Khan F, Yang M-H. Intriguing properties of vision transformers. Adv Neural Inf Process Syst 2021;34:23296–308.
[23] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proc IEEE/CVF Int Conf Comput Vis (ICCV); 2021. p. 10012–22.
[24] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc Eur Conf Comput Vis (ECCV); 2018. p. 801–18.
[25] Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2021. p. 2613–22.
[26] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inf Process Syst 2017;30.
[27] Hinton JP, et al. A method to reuse archived h&e stained histology slides for a multiplex protein biomarker analysis. Methods Protoc 2019;2:86.
[28] Li X, et al. Unsupervised content-preserving transformation for optical microscopy. Light: Sci Appl 2021;10:44.
[29] de Bel T, Bokhorst J-M, van der Laak J, Litjens G. Residual cyclegan for robust domain transformation of histopathological tissue slides. Med Image Anal 2021;70:102004.
[30] Ruder S. An overview of gradient descent optimization algorithms. Preprint. arXiv: 1609.04747, 2016.
[31] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2016. p. 770–8.
[32] Rivenson Y, Liu T, Wei Z, Zhang Y, de Haan K, Ozcan A. Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light: Sci Appl 2019;8:1–11.
[33] Zhang Y, de Haan K, Rivenson Y, Li J, Delis A, Ozcan A. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. Light: Sci Appl 2020;9:78.
[34] de Haan K, Zhang Y, Zuckerman JE, Liu T, Sisk AE, Diaz MF, et al. Deep learning-based transformation of h&e stained tissues into special stains. Nat Commun 2021;12:1–13.
[35] Park Y, Depeursinge C, Popescu G. Quantitative phase imaging in biomedicine. Nat Photonics 2018;12:578–89.
[36] Li Z, Sun J, Fan Y, Jin Y, Shen Q, Trusiak M, et al. Deep learning assisted variational Hilbert quantitative phase imaging. Opto-Electron Sci 2023;2:220023.
[37] Lu L, Li J, Shu Y, Sun J, Zhou J, Lam EY, et al. Hybrid brightfield and darkfield transport of intensity approach for high-throughput quantitative phase microscopy. Adv Photon 2022;4:056002.
[38] Zhou S, Li J, Sun J, Zhou N, Ullah H, Bai Z, et al. Transport-of-intensity Fourier ptychographic diffraction tomography: defying the matched illumination condition. Optica 2022;9:1362–73.
[39] Liu R, Sun Y, Zhu J, Tian L, Kamilov US. Recovery of continuous 3d refractive index maps from discrete intensity-only measurements using neural fields. Nat Mach Intell 2022;4:781–91.
[40] Park J, Shin S-J, Shin J, Lee AJ, Lee M, Lee MJ, et al. Quantification of structural heterogeneity in h&e stained clear cell renal cell carcinoma using refractive index tomography. Biomed Opt Express 2023;14:1071–81.
[41] Li J, Zhou N, Sun J, Zhou S, Bai Z, Lu L, et al. Transport of intensity diffraction tomography with non-interferometric synthetic aperture for three-dimensional label-free microscopy. Light: Sci Appl 2022;11:154.
[42] Zhou S, Li J, Sun J, Zhou N, Chen Q, Zuo C. Accelerated Fourier ptychographic diffraction tomography with sparse annular led illuminations. J Biophotonics 2022;15:e202100272.
[43] Bucher M, Vu T-H, Cord M, Pérez P. Zero-shot semantic segmentation. Adv Neural Inf Process Syst 2019;32.