

RESEARCH

Open Access



A human-in-the-loop method for pulmonary nodule detection in CT scans

Qingjie Zeng^{1†} , Yutong Xie^{2†}, Zilin Lu¹ and Yong Xia^{1*}

Abstract

Automated pulmonary nodule detection using computed tomography scans is vital in the early diagnosis of lung cancer. Although extensive well-performed methods have been proposed for this task, they suffer from the domain shift issue between training and test images. Unsupervised domain adaptation (UDA) methods provide a promising means to mitigate the domain variance; however, their performance is still limited since no target domain supervision is introduced. To make the pulmonary nodule detection algorithm more applicable in clinical practice and further boost the performance across domains, we propose a human-in-the-loop method in a semi-supervised fashion to enhance the model generalization ability when transferred from source domain to target domain. Specifically, we first train a detector model on source domain, and then the pre-trained detector is utilized with our proposed uncertainty-guided sample selection scheme (USSS) to find a few target domain samples worth annotating most and obtain their human annotations. Finally, the annotated and the rest unlabeled target domain samples are used together to refine the pre-trained model via our proposed zoom-in and zoom-out constraint (ZZC) strategy. We evaluate our method on the Nodule Analysis 2016 (LUNA16) and TianChi datasets. Experimental results show that our method surpasses recent competitive methods on source domain and also achieves surprising performance on target domain.

Keywords: Pulmonary nodule detection, Human-in-the-loop, Semi-supervised learning, Domain adaptation

1 Introduction

Lung cancer is one of the major causes of cancer death over the world [1]. The average five-year survival rate of lung cancer patients with metastatic disease is approximately 6%. If an early stage diagnosis and timely treatment are implemented, the survival rate could be improved to 60% [2]. Malignant pulmonary nodules have a high probability of evolving into malignant tumors, and thus early detection and diagnosis of pulmonary nodules are vital in patient care. Generally, pulmonary nodules are isolated and spherical on chest computed tomography (CT) scans, which usually refer to spots with diameters of less than

30 mm. To examine patients, radiologists inspect CT scans on slice-by-slice basis, which is time-consuming and prone to operator bias. Computer-aided pulmonary nodule detection avoids many of these problems and has been studied extensively [3–9].

The fast and successful development [8, 10–13] of convolutional neural networks (CNNs) boosts CNN-based pulmonary nodule detection methods, in which nodule candidate detection and false positive reduction are usually two essential steps. In recent studies, Mei et al. [8] proposed a slice-aware grouped non-local module to capture long-range dependencies in the stage of nodule candidate detection, and then employed a 3D CNN for false positive reduction. Luo et al. [7] introduced an end-to-end and anchor-free network that uses sphere representation and center points matching to detect pulmonary nodules. Zhou et al. [12] established a 2.5D multi-stage cascaded network for efficient and robust nodule detec-

* Correspondence: yxia@nwpu.edu.cn

[†]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
Full list of author information is available at the end of the article [†]Equal contributors

tion. Inspired by clinical examination, Zheng et al. [14] utilized maximum intensity projection images with different thicknesses as input, which effectively increased the sensitivity of computer-aided diagnosis (CAD) system. Considering that nodules have various shapes and sizes, Guo et al. [15] developed an aggregation network and a self-calibrated strategy to help the detector discover more low-resolution details contained in the high-level features. Despite their promising performance, these methods suffer from domain shift between training and test images. Since the quality of medical images depends on many factors such as the imaging equipment, scanning protocol, and expertise of the operator (see Fig. 1), the image quality may vary between training and test datasets, causing the performance of a well-trained model to drop unexpectedly on an unseen test set.

Recently, domain adaptation (DA) methods have drawn increasing attention to address the aforementioned troublesome issues [16]. DA mainly focuses on data transformation [17], feature alignment [18], soft label [19] and ensemble learning [20] to reduce the distribution discrepancy across domains. For instance, completing image-to-image translation by a generative adversarial network (GAN) [21], establishing auxiliary tasks to learn the prior knowledge and domain-invariant information [22], and transferring knowledge in a self-supervised manner [23]. Although these strategies successfully narrow down domain shift between the training and test datasets, their performance is still limited by three drawbacks: (1) proxy tasks are required for model knowledge adaptation; (2) the usage of GAN is still troublesome, since the quality and structures of the generated images cannot be guaranteed; and (3) source and target datasets need to be used at the same time.

Human-in-the-loop (HITL) has been mooted to integrate human prior knowledge into the procedure of machine training [24, 25]. The purpose of HITL is to train a robust model, which can reduce the sparse but fatal mistakes, especially in the risk-unbearable fields, i.e., clinical diagnosis and autonomous driving. Concretely, in the aspect of data annotation, HITL aims to minimize the human labor while maintaining the model performance, when raw data continuously accumulate [26].

In this paper, we propose a human-in-the-loop method for pulmonary nodule detection in a semi-supervised fashion to enhance the model generalization ability when transferred from source domain to target domain. To be specific, an uncertainty-guided sample selection scheme (USSS) is proposed to select a few target domain samples worth annotating via aleatoric and epistemic uncertainty. Beyond that, a novel constraint strategy called zoom-in and zoom-out constraint (ZZC) is constructed. ZZC uses these annotated target domain samples and rest unlabeled target domain samples to learn scale-invariant

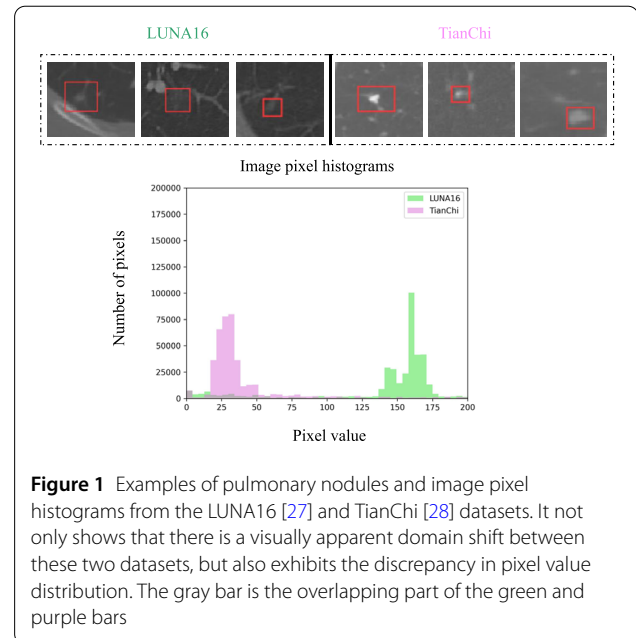


Figure 1 Examples of pulmonary nodules and image pixel histograms from the LUNA16 [27] and TianChi [28] datasets. It not only shows that there is a visually apparent domain shift between these two datasets, but also exhibits the discrepancy in pixel value distribution. The gray bar is the overlapping part of the green and purple bars

but consistent information for refining the pre-trained source domain model in a semi-supervised fashion. The motivation is that the relative spatial location of nodules does not change even if the shape varies. As indicated by our results, the proposed method can achieve state-of-the-art pulmonary nodule detection performance in both source and target domains.

The main contributions of this work are three-fold: (1) we propose a human-in-the-loop method in a semi-supervised fashion for pulmonary nodule detection in chest CT images, which has an improved generalization ability and is more suitable for clinical practice; (2) we design an uncertainty estimation scheme to effectively discover target domain samples with high uncertainty, which are worth annotating; and (3) we propose a novel constraint to explore scale-invariant features, which help the model to learn consistent target domain information.

2 Related work

2.1 Domain adaptation

Domain adaptation (DA) has been widely studied to address the domain shift issue between source domain and target domain [16, 29]. Specially, another topic called unsupervised domain adaptation (UDA) is developed when the labeled target domain samples are not available. Generally, UDA reduces the distribution discrepancy by means of image transformation, feature alignment, ensemble learning and more. Yan et al. [30] developed a Unet-GAN, which first trained a Unet [31] on source domain, then transformed the style of target data similar to source domain, and finally conducted segmentation task on the transformed data. Similarly, Wolterink et al. [17] utilized Cycle-GAN [21] to perform image-to-image translation,

and then minimized the difference between the original and reconstructed images. Liu et al. [23] proposed a source-free domain adaptation method, which used knowledge transferring in self-supervised manner, to solve the problem where only a well-trained source model and an unlabeled target dataset were available for optimization.

Although UDA methods have achieved good results under a dedicated design, the performance gain is still limited since the quality of generated images cannot be guaranteed and specified domain knowledge transferring tasks are required.

2.2 Semi-supervised learning

Considering the time cost and human labor of annotation, semi-supervised learning (SSL) has gained increasing popularity in medical image analysis [32, 33]. Existing methods generally focus on pseudo labeling [34–36], consistency regularization [37–39] and adversarial training [40]. Specifically, pseudo labeling methods first train a model on labeled data, then generate pseudo labels for the rest unlabeled data, and finally select trustworthy pseudo labels for re-training until convergence. Yu et al. [41] constructed an uncertainty-aware mean teacher framework, where the teacher network is the historical ensemble of the student network and generates pseudo labels to guide the student network learning. Consistency regularization encourages the predicted outputs to be consistent, under the scenarios of adding noise, utilizing differentiated predictors and data transformation. For example, Luo et al. [42] proposed a dual-task regularization strategy to help model learn task-level consistency. Wu et al. [43] developed two decoders with different structures and output level consistency. Adversarial training aims to train a robust model with targeted perturbations. Miyato et al. [40] introduced a virtual training strategy, which injected random noise first, then obtained the gradient in the direction of noise through loss back-propagation, and subsequently used the normalized gradient as the final noise for model training.

The above SSL methods mainly focus on generating reliable pseudo labels or consistent outputs to help the model learn from unlabeled samples, but in the task of pulmonary nodule detection, nodules are typically small and diverse, making it difficult to obtain reliable pseudo labels, and simple consistency regularization has little effect on unlabeled samples. Moreover, these SSL methods are performed on the same domain, which may not help to mitigate the domain gap in the face of domain shift.

2.3 Human-in-the-loop AI

Human-in-the-loop (HITL) aims to complete tasks under human-machine collaboration [24, 25]. In medical image analysis, especially in medical image segmentation, dense annotations are required to train a model. However,

this procedure is complicated and demands professional knowledge. Beyond that, it is almost impossible to annotate all data since the growth rate of new tasks has far exceeded the acceptable labeling cost [44]. For instance, it is unlikely to label abundant different domain samples when facing numerous new scenes. To address the above issues, Wang et al. [45] developed a two-stage system, which focused on the sparse but fatal cases. This system tried to leverage failure samples to calibrate pre-trained models with little annotation cost. Ravanbakhsh et al. [46] utilized conditional GAN(cGAN) [47] to complete magnetic resonance image (MRI) segmentation. Specifically, generator and discriminator were first trained with a few labeled samples, then generator was utilized to segment unlabeled samples and discriminator was used to distinguish whether the input was a hard case, and finally, easy cases with pseudo labels and hard cases with human annotation were collected to retrain the model. On the task of landmark detection, Quan et al. [48] selected the templates worth annotating by estimating the representation similarity within each image and the template group.

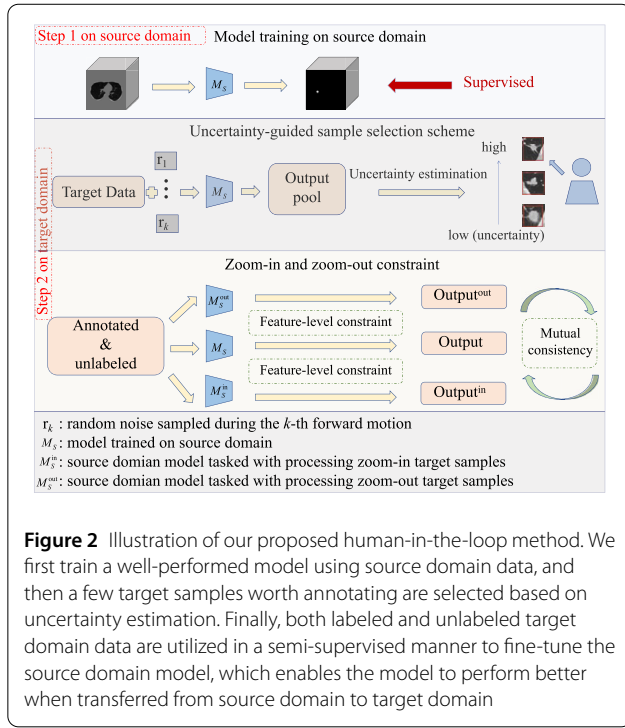
Similarly, in this work, from the aspect of using limited human labeling cost to enhance model generalization ability, we propose a pulmonary nodule detection method in the scenario of semi-supervised learning and human-in-the-loop, with the aim of mitigating the performance degradation caused by domain shift.

3 Methods

3.1 Preliminary

Let $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and $D_T = \{D_{UT}, D_{LT}\}$ denote the source domain dataset and target domain dataset, respectively. Here N is the number of samples, $D_{UT} = \{(x_i^{ut})\}_{i=1}^{N_{ut}}$ contains N_{ut} unlabeled samples, and $D_{LT} = \{(x_i^{lt}, y_i^{lt})\}_{i=1}^{N_{lt}}$ contains N_{lt} annotated samples collected by our uncertainty-guided sample selection scheme.

Based on the above definition, the workflow of our human-in-the-loop method for pulmonary nodule detection is depicted in Fig. 2, which mainly contains three parts. First, we train a model using D_S in a regular routine, and make sure that the model performs well on the source domain. Nevertheless, due to the domain shift issue, the pre-trained model may suffer from performance degradation when transferred to a different domain. To bridge the domain gap in a feasible way, we propose the uncertainty-guided sample selection scheme to find a few target domain samples worth annotating and create a labeled target domain dataset, termed D_{LT} . Finally, we utilize both D_{LT} and D_{UT} in a semi-supervised manner with the zoom-in and zoom-out constraint to fine-tune the model pre-trained on D_S . The goal of this task is to enhance the model generalization ability when transferred from source domain to target domain.



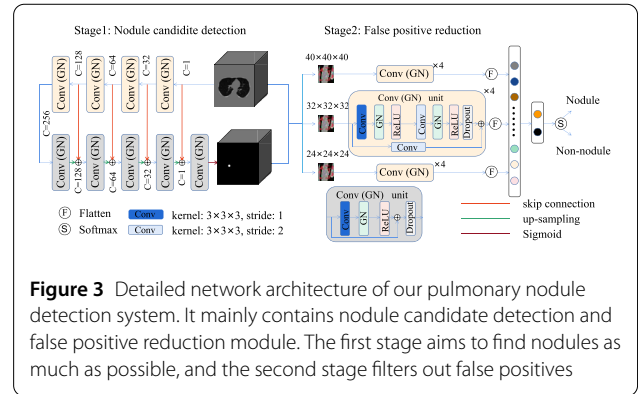
3.2 Training on source domain

3.2.1 Nodule candidate detection

Pulmonary nodule detection commonly has two steps: nodule candidate detection and false-positive reduction, as shown in Fig. 3. In this paper, we construct a U-Net-like [31] architecture for nodule candidate detection. The encoder consists of four repeated 3D residual blocks [49], each containing two convolutional layers. Each layer is followed by the group normalization (GN) [50] and ReLU activation. In each down-sampling stage, we use the $3 \times 3 \times 3$ convolution with two strides to halve the resolutions of input feature maps and double the number of channels. The decoder is symmetrically constructed and the skip connection is utilized to link low-level and high-level feature maps. At each up-sampling stage, features are summed first and then refined by a residual block.

3.2.2 False positive reduction

The proposed pulmonary nodule detection method follows a two-stage pipeline. Specifically, the nodule candidate detection module plays the role of finding all potential nodules through segmentation. Based on the segmentation results, we can acquire the minimum and maximum values of the x -axis, y -axis and z -axis for each candidate. Therefore, the center coordinates of each candidate can be calculated. Then, the candidates will be cropped according to the center coordinates. So far, these candidates include not only the true pulmonary nodules, but also the false positive background. To this end, the false positive reduction module is further employed with the aim of filter-



ing out the false positives. This is achieved through binary classification (nodule or non-nodule). Thus, under the assistance of the false positive reduction module, the false positive rate can be reduced. Specifically, the false positive reduction module is constructed to estimate the probability of being a nodule from multi-scale, which contains a three-branch feature encoder E and a classifier CLS composed of two fully connected layers. The architecture of each single branch in E is the same as the encoder of the nodule candidate detection module except for different input sizes. Considering that pulmonary nodules have complicated contextual characteristics and vary significantly in size, solidity and sphericity, we crop cubic patches with shapes of $40 \times 40 \times 40$, $32 \times 32 \times 32$ and $24 \times 24 \times 24$ based on the detection results as inputs, and concatenate the final flattened features to obtain the nodule prediction results.

3.2.3 Loss function

For nodule candidate detection, we use the sum of the Dice loss and focal loss, since the volume of a pulmonary nodule may be very small. In the false positive reduction, we simply use the cross-entropy loss to optimize the classifier.

$$L_{\text{Dice}}(p, y) = 1 - \frac{2 \sum_{i=1}^{N_p} p_i y_i}{\sum_{i=1}^{N_p} p_i + \sum_{i=1}^{N_p} y_i}, \quad (1)$$

$$L_{\text{focal}}(p, y) = -\frac{1}{N_p} \sum_{i=1}^{N_p} (1 - p_i)^{\gamma} y_i \ln p_i - \frac{1}{N_p} \sum_{i=1}^{N_p} p_i^{\gamma} (1 - y_i) \ln (1 - p_i), \quad (2)$$

$$L_{\text{det}}(p, y) = L_{\text{Dice}}(p, y) + L_{\text{focal}}(p, y), \quad (3)$$

where L_{det} is the loss for detector, and p_i and y_i are the predicted probability and ground truth of the i -th voxel, respectively. N_p is the number of voxels and γ is a hyper-parameter.

3.3 Uncertainty-guided sample selection scheme

In clinical practice, domain shift between training data and test data is inevitable considering the factors of imaging quality, such as scanning protocol, imaging equipment and expertise of the operator. Although the UDA methods have been widely explored to address this issue, the performance is still limited since no target label information is introduced during the model fine-tuning procedure.

Empirically, inherent random effect and ambiguity in data, and Monte Carlo dropout indicate aleatoric uncertainty (AU) and epistemic uncertainty (EU), respectively. The former shows data diversity and the latter represents model diversity [51]. Following this observation, we add different random perturbations during multiple forward passes with a certain dropout rate, which incorporates AU and EU to assess the sample uncertainty. Unlike traditional uncertainty estimation methods that add noise singly (aleatoric uncertainty), here we additionally perform multi-times forward, which incorporates epistemic uncertainty into the estimation since the dropout layer is included in the model. Therefore, an uncertainty-guided sample selection scheme (USSS) is proposed to select a few target samples worth annotating most based on the model pre-trained on D_S . Specifically, for each sample x_i^{ut} in original D_T , different random noise r_i^k is added for K iterations (the noise is first sampled from standard normal distribution $N \sim (0,1)$, whose value is then truncated within -0.1 to 0.1), thus, K predicted soft masks $\{m_i^k\}_{k=1}^K$ are generated for x_i^{ut} , and the uncertainty score of x_i^{ut} is calculated using Eq. (4):

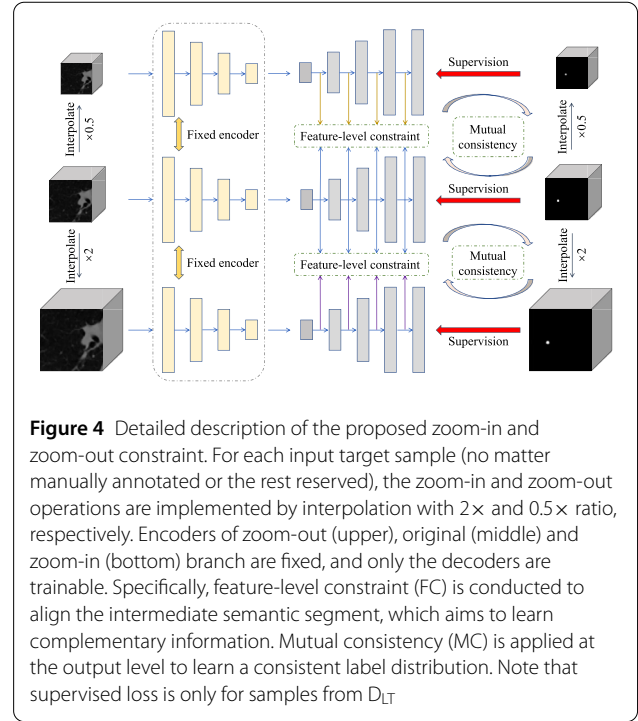
$$U_i = 1 - \frac{\sum_{k=1}^K \mathbb{1}[m_i^k \geq \tau]}{\sum_{k=1}^K \mathbb{1}[m_i^k \geq \tau]}, \quad (4)$$

where τ is a hyper-parameter used to obtain binary masks. K is the iteration number. The higher U_i , the more unstable the segmentation is.

When uncertainty of all samples in original unlabeled target domain dataset D_T is calculated, we collect a few samples with high uncertainty score to annotate and form a labeled target domain dataset termed D_{LT} .

3.4 Zoom-in and zoom-out constraint

Beyond the collected D_{LT} , there are a large number of unlabeled available samples, which form D_{UT} . Inspired from previous works in which multi-level contextual information indeed boosts the performance of pulmonary nodule detection [4, 14], we further propose a ZZC strategy to optimize the pre-trained model in a semi-supervised manner, with the aim of adapting the model to unseen target domain perfectly. As shown in Fig. 4, ZZC mainly contains feature-level constraint (FC) and mutual consistency



(MC), which are designed for intermediate segment alignment and predicted result consistency, respectively.

3.4.1 Feature-level Constraint (FC)

Compared to the features of the middle branch, features of the zoom-in and zoom-out branch are $2\times$ magnification and $0.5\times$ scaling along depth, height and width axis, respectively. We argue that semantic information at each level should be consistent even if the input scale varies, since the relative spatial location of a nodule is not changed. Formally, the constraint between upper and middle branch and the constraint between bottom and middle branch are formulated as below:

$$L_{\text{FC}} = L_{\text{FC}}^{u \leftrightarrow m} + L_{\text{FC}}^{b \leftrightarrow m}, \quad (5)$$

$$L_{\text{FC}}^{u \leftrightarrow m} = \frac{1}{I} \frac{1}{CDHW} \sum_{i=1}^I \|\phi(\sigma(f_i^u)) - \sigma(f_i^m)\|_2^2, \quad (6)$$

$$L_{\text{FC}}^{b \leftrightarrow m} = \frac{1}{I} \frac{1}{CDHW} \sum_{i=1}^I \|\psi(\sigma(f_i^b)) - \sigma(f_i^m)\|_2^2, \quad (7)$$

where f_i^u , f_i^m and f_i^b mean the i -th level feature from upper, middle and bottom branch, respectively. σ denotes sigmoid function and ϕ stands for up-sampling. ψ is the max-pooling function and I is the number of scales. C , D , H and W represent the channel, depth, height and width of the feature map, respectively. L_{FC} is the final loss of FC.

To sum up, L_{FC} assists decoders in capturing scale-invariant and scale-complementary semantic information as multi-level multi-view features are interacted.

3.4.2 Mutual Consistency (MC)

In addition to the feature-level consistency learning, we also introduce the output level consistency for the pre-trained model fine-tuning. Actually, MC works in the manner of pseudo-labeling, where each branch generates pseudo mask for the other. Moreover, following Ref. [52], we utilize a sharpening function to transform the predicted soft pseudo mask into sharpened pseudo mask, which makes contribution to entropy regularization. The sharpening function $\xi(\cdot)$ is formulated as

$$\xi(p_i) = \frac{p_i^{1/T}}{p_i^{1/T} + (1 - p_i)^{1/T}}, \quad (8)$$

where T is the sharpening parameter, and p_i is the predicted probability of the i -th voxel. The smaller the value T , the closer the sharpened pseudo mask to a binary mask.

Similar to FC, the loss between upper and middle branch and loss between bottom and middle branch are formulated as below:

$$L_{MC} = L_{MC}^{u \leftrightarrow m} + L_{MC}^{b \leftrightarrow m}, \quad (9)$$

$$L_{MC}^{u \leftrightarrow m} = L_{\det}(\phi(p^u), \xi(p^m)), \quad (10)$$

$$L_{MC}^{b \leftrightarrow m} = L_{\det}(\psi(p^b), \xi(p^m)), \quad (11)$$

where p^u , p^m and p^b are the predicted soft pseudo masks from the upper, middle and bottom branch, respectively. L_{MC} is the final loss of MC.

Based on the above procedure, the fine-tuning of nodule candidate detection has finished. As for the false positive reduction fine-tuning, feature encoder E is fixed and only the classifier CLS composed of two fully connected layers is trainable. Similarly, there are labeled and unlabeled pulmonary nodules, which are from annotated samples and detected by the fine-tuned detector, respectively. As for the detected unlabeled pulmonary nodules, we simply utilize the pseudo-labeling strategy with the threshold η to filter out untrustworthy pseudo labels (nodule or non-nodule), and iterate until convergence.

4 Experimental results

4.1 Dataset and evaluation metrics

We evaluate our method on the LUNA16 dataset [27] and the TianChi dataset [28]. LUNA16 contains 888 low-dose CT scans along with 1186 pulmonary nodules, which is composed of 10 subsets. The diameter of these pulmonary nodules is 3~30 mm. The TianChi dataset contains 1000 low-dose CT scans (600 for training and 400 for testing) with 1244 pulmonary nodules, and the diameter of these

pulmonary nodules is 3~30 mm except one with 2.9 mm. For both two datasets, the center coordinates and diameters of the pulmonary nodules are given as annotation information. In main experiments, we consider the LUNA16 dataset as the source domain and the TianChi dataset as the target domain.

Evaluation metric for all experiments is the free-response receiver operating characteristic (FROC) [54] under seven pre-defined false positives per scan (FP/scan) rates: 0.125, 0.25, 0.5, 1, 2, 4, and 8. The averaged sensitivity of these seven pre-defined FP/scan is referred to as competition performance metric (CPM) [55]. The FROC curve shows the relevance between sensitivity and FPs/scan.

4.2 Implementation details

In data pre-processing, following Refs. [7, 9, 56], we first normalize all chest CT scans to a unified voxel size of $1 \times 1 \times 1 \text{ mm}^3$ to reduce the influence of spatial variation across different samples. Following Refs. [12, 14, 15], we employ a lung parenchyma segmentation scheme to remove irrelevant anatomical structures. This scheme contains four steps: (1) using the mean value of the whole slices as a threshold to roughly divide each case into the lung and background; (2) keeping the largest two connected regions as pulmonary regions; (3) using morphological dilation and closing to fill holes to ensure that the nodules adhering to the lung wall are contained by the lung parenchyma mask; and (4) applying the mask to the scan to remove irrelevant anatomy.

For training details, we extract partly overlapped 3D patches of size $48 \times 48 \times 48$ on the re-sliced CT scans with a stride of 40. To train the nodule candidate detection module, we generate pseudo nodule segmentation masks through two steps: (1) defining a sphere with the same diameter as the nodule and the center of the sphere is located in the center of the pulmonary nodule; (2) using the Butterworth function to set voxel value inside the sphere and set 0 to the outside. Beyond that, we utilize data augmentation including random flip and shifting (-10 to 10 voxels) to enlarge the number of positive samples (cubic patches containing nodules), which balances the negative samples (cubic patches without nodules). An Adam optimizer with an initial learning rate of 1×10^{-4} and 1×10^{-6} is used to optimize the nodule candidate detection module and false positive reduction module, respectively. For source domain training, the batch size and hyper-parameter γ are empirically set to 32 and 2.5, respectively. For target domain fine-tuning, the batch size, hyper-parameter τ , iteration number K , sharpening parameter T and pseudo-labeling threshold η are empirically set to 8, 0.35, 3, 0.1 and 0.70, respectively. Moreover, we evaluate our method using 10-fold cross-validation and test set validation on the LUNA16 and the TianChi datasets, respectively. All experiments were conducted on two NVIDIA Gefore RTX 3080Ti GPUs.

Table 1 Performance comparison with other competitive pulmonary nodule detection methods on the LUNA16 dataset [27]. FPR means false positive reduction; CPM is competition performance metric. It is the average sensitivity of seven pre-defined FPs/scan. FPs mean false positives per scan. Best results are in bold

Method	FPR	FPs/scan							CPM
		0.125	0.25	0.5	1	2	4	8	
CPM-Net [9]	w/o	0.723	0.838	0.887	0.911	0.928	0.934	0.948	0.881
SCPM-Net [7]	w/o	0.743	0.829	0.889	0.922	0.939	0.958	0.964	0.892
MIP [14]	w	0.876	0.899	0.912	0.927	0.942	0.948	0.953	0.922
TSCNN [53]	w	0.848	0.900	0.925	0.936	0.949	0.957	0.960	0.925
DS-CMSF [12]	w	0.742	0.841	0.899	0.925	0.944	0.955	0.960	0.895
MSANet [15]	w	0.832	0.886	0.928	0.937	0.946	0.952	0.959	0.920
Ours	w	0.878	0.916	0.933	0.954	0.962	0.968	0.970	0.940

Table 2 Performance comparison with UDA, SSL and USS methods on the target domain TianChi dataset [28]. “UDA”, “SSL” and “USS” are unsupervised domain adaptation, semi-supervised learning and unsupervised semantic segmentation, respectively. Note that all SSL methods using the same 5% labeled target domain samples, which are selected by our proposed USSS

Method	Type	FPs/scan							CPM
		0.125	0.25	0.5	1	2	4	8	
Baseline (source)	–	0.672	0.771	0.815	0.832	0.855	0.878	0.882	0.815
Baseline (target)	–	0.771	0.815	0.846	0.902	0.926	0.941	0.953	0.879
LUSeg [57]	USS	0.657	0.763	0.801	0.822	0.847	0.860	0.878	0.804
UNet-GAN [30]	UDA	0.802	0.856	0.902	0.908	0.912	0.916	0.918	0.888
SFDA [23]	UDA	0.846	0.874	0.900	0.911	0.918	0.922	0.923	0.899
ConstraintAdap [22]	UDA	0.835	0.862	0.905	0.917	0.923	0.936	0.938	0.902
UA-MT [41]	SSL	0.835	0.862	0.905	0.927	0.936	0.938	0.938	0.906
CPS [34]	SSL	0.843	0.886	0.912	0.936	0.938	0.941	0.941	0.914
ST++ [58]	SSL	0.851	0.904	0.909	0.932	0.941	0.943	0.946	0.918
Ours	SSL	0.871	0.916	0.928	0.943	0.948	0.952	0.958	0.931

4.3 Comparison with other detection methods on source domain LUNA16 dataset

Table 1 shows the performance comparisons with other competitive pulmonary nodule detection methods on the LUNA16 dataset. All results are obtained from the original papers and based on the same 10-fold cross-validation. Generally, these methods can be divided into one-stage and two-stage approaches, depending on whether a false positive reduction module is used for further classification. Based on the results, we can find that two-stage methods (MIP [14], TSCNN [53], DS-CMSF [12], and MSANet [15]) are better than one-stage methods (CPM-Net [9] and SCPM-Net [7]), with approximately 1.4%~4.4% performance gain, manifesting the necessity of utilizing false positive reduction. Beyond that, our method outperforms other existing two-stage methods by 2.0% compared with MSANet. The main reason is that we focus on detailed nodule representation with multi-scale features on over-sampled balanced data, and finally achieve the highest CPM score of 0.940. Moreover, it is noticed that the FPs/scan rates from 1 to 4 are the most clinical concerns [59]. Our method yields sensitivity of 0.954, 0.962

and 0.968 under FPs/scan rates of 1, 2, and 4, respectively, which indicates the promising potential of our detection method on the source domain data.

4.4 Comparison with UDA and SSL methods on target domain TianChi dataset

To validate the effectiveness of our human-in-the-loop method, we compare it with the recently proposed UDA and SSL methods, respectively, which are shown in Table 2. Here baseline (source) means directly applying the model trained on the LUNA16 dataset to the TianChi dataset, which results in the lowest performance and further indicates the domain shift between the two datasets. Baseline (target) means that only the selected 5% target domain samples are used to fine-tune the pre-trained model, which achieves 6.4% performance gain compared to baseline (source). This result demonstrates that providing a small amount of target domain supervision can contribute to better accuracy.

Comparisons with the UDA methods. Three UDA methods are utilized for comparison, which are UNet-GAN [30], SFDA [23] and ConstraintAdap [22]. Based on the results

in Table 2, UNet-GAN, SFDA and ConstraintAdap obtain the CPM score of 0.888, 0.899 and 0.902, which is 7.3%, 8.4% and 8.7% higher than baseline (source), respectively. This result demonstrates their effectiveness in mitigating domain shift. However, it is interesting to find that all UDA methods have lower CPM score than SSL methods with more than 1.6% performance gap. This indicates that supplying a small amount of target domain supervision would greatly boost the performance.

Comparisons with SSL methods. Three recently proposed advanced SSL methods are listed for comparison: UA-MT [41], CPS [34], and ST++ [58]. To make fair comparison, all these SSL methods utilize the same 5% labeled target domain samples collected by our proposed USSS. From the results in Table 2, we can see that ST++ and CPS achieve similar performance and surpass UA-MT in CPM score by 0.8%~1.2%. Beyond that, CPS and ST++ surpass baseline (source) by 9.9% and 10.3%, respectively. This suggests that fine-tuning the pre-trained model in the SSL method can visibly help the model learn target domain information. More importantly, our method obtains 93.1% CPM score and has approximately 2% improvement compared to the other SSL methods, which demonstrates the ability of the proposed ZZC strategy to extract target domain information from unlabeled samples.

Additionally, we also compare with an unsupervised semantic segmentation method LUSeg [57] with a few annotation target domain samples. Specifically, to make LUSeg more suitable for the dense prediction task, i.e., segmentation, we re-implement LUSeg by first pre-training on all target domain data in a masked image modeling fashion, instead of contrast-based pre-training. Then, LUSeg is fine-tuned using 5% target domain annotations selected by our proposed USSS. From Table 2, we can find our method surpasses LUSeg by 12.7%. We believe there are two reasons for this: (1) although LUSeg can learn target semantics by pre-defined proxy tasks, the learned features are optimal to the pertinent high-level segmentation task, while task-specific semantics are only acquired from the limited 5% labeled target domain samples, resulting in relatively poor performance; (2) our method is first pre-trained using labeled source domain samples, which enables our method to perform well on the current pulmonary nodule detection task. Although there is a domain gap between the source and target domains, the domain shift problem can be mitigated by jointly leveraging a few target samples worth annotating and a large set of unlabeled target samples in a semi-supervised manner.

To further demonstrate our method, we randomly select seven samples from source domain LUNA16 and target domain TianChi dataset, respectively. As shown in Fig. 5, the first row “source domain” shows that our model can accurately find nodules with high certainty when training and testing are performed on the source domain LUNA16

dataset. The rest four rows are detection results when model transferred from source domain to target domain. Here we compare our method to one best UDA method (ConstraintAdap) and two best SSL methods (CPS, ST++) according to the results in Table 2. From the visualization results, we can draw three conclusions: (1) ConstraintAdap can capture easy samples, i.e., solitary and relatively large nodules, but failed when faced with hard samples; (2) although CPS and ST++ have fewer true negatives compared to ConstraintAdap, they are still sensitive to small nodules or nodules attached to tissues (see last three columns in Fig. 5), resulting in mislocalization and low probability of nodule detection; (3) in contrast, our method achieves better results with accurate location and high classification score, further indicating that the proposed ZZC strategy can effectively explore scale-invariant but consistent features from both feature and output levels.

4.5 Ablation study

4.5.1 Number of human annotations

As depicted in Fig. 6, six ratios of annotated target domain samples are discussed, which are 1%, 2%, 5%, 10%, 20% and 50%. Based on the results, three conclusions are drawn as follows: (1) the performance of these methods are gradually increasing when the annotation ratio ranges from 1% to 10%, and results are almost saturated when the ratio is greater 10%; (2) our method significantly surpasses other competitive SSL methods when the ratio is low, i.e., there is about 2.5% performance gain compared to ST++ when only 1% target domain samples are labeled. This phenomenon indicates that our method is more suitable to discover target domain information in a semi-supervised manner with limited supervision; and (3) in the scenario of annotating abundant target domain samples, i.e., 50% ratio, approximately 1% performance gain is achieved when compared to ST++, which demonstrates our method is more capable of mining the information of unlabeled target domain samples.

4.5.2 Comparisons with different uncertainty estimation methods

We also conduct experiments to validate the superiority of our proposed uncertain estimation scheme. As presented in Table 3, four strategies are compared, which are Random (randomly select target domain samples to annotate), Entropy, Noise (AU) and Dropout (EU) [60], in the case of only utilizing the selected 5% target domain samples to refine the model. From the results, we can see that the baseline results in the lowest performance. Noise and Dropout outperform Entropy and Random, which indicating that AU and EU are more suitable for estimating instance uncertainty. The performance gain between USSS

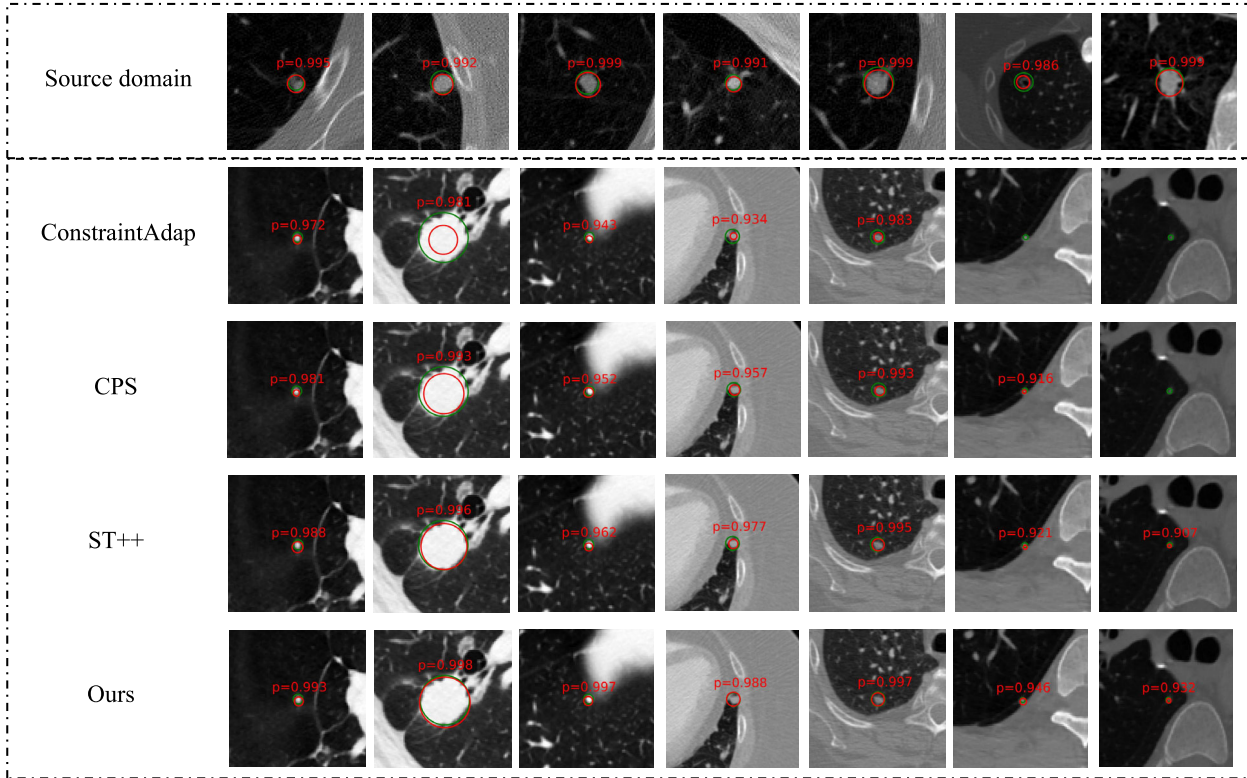


Figure 5 Visualization of pulmonary nodule detection results. The first row “source domain” means the model is trained and tested on the source domain LUNA16 dataset. The remaining four rows perform comparisons with the UDA and SSL methods when the pre-trained model is transferred from source domain to target domain. The red circle and “p” denote the location and probability of the detected nodule, respectively. The green circle is the ground truth

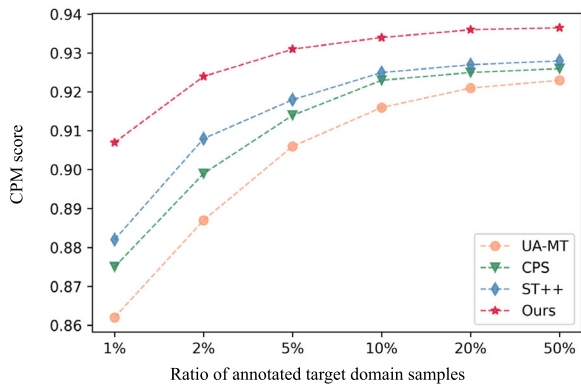


Figure 6 Comparison with other SSL methods with different numbers of annotated target domain samples

and Noise/Dropout demonstrates the effectiveness of incorporating AU and EU, which significantly boosts the performance when only limited target domain samples are worth annotating and delineated.

Table 3 Performance comparison with other uncertainty estimation methods on the target domain TianChi dataset, in the case of annotating 5% target domain samples

Method	FPs/scan							CPM
	0.125	0.25	0.5	1	2	4	8	
Baseline	0.672	0.771	0.815	0.832	0.855	0.878	0.882	0.815
Random	0.692	0.803	0.824	0.857	0.886	0.902	0.913	0.840
Entropy	0.742	0.807	0.838	0.878	0.903	0.927	0.931	0.861
Noise	0.753	0.808	0.836	0.880	0.908	0.927	0.930	0.863
Dropout	0.763	0.814	0.842	0.889	0.912	0.930	0.933	0.869
USSS	0.771	0.815	0.846	0.902	0.926	0.941	0.953	0.879

4.5.3 Effectiveness of each module

From Table 4, we can find that (1) just introducing USSS can produce approximately 6% CPM improvement over baseline. Here, B+USSS means using only selected 5% samples from the target domain to fine-tune the pre-trained source domain model. The result shows that supplying a small amount of target domain supervision is quite useful to bridge the domain gap; (2) in the absence of annotated target samples, the use of ZZC strategy effectively

Table 4 Quantitative ablation study of each module on target domain TianChi dataset

Method	FPs/scan							CPM
	0.125	0.25	0.5	1	2	4	8	
Baseline(B)	0.672	0.771	0.815	0.832	0.855	0.878	0.882	0.815
B+USSS	0.771	0.815	0.846	0.902	0.926	0.941	0.953	0.879
B+ZZC	0.835	0.862	0.905	0.917	0.936	0.938	0.940	0.905
B+USSS+ZZC	0.871	0.916	0.928	0.943	0.948	0.952	0.958	0.931

Table 5 CPM score with different τ on target domain TianChi

τ	0.15	0.25	0.35	0.45	0.55	0.65
CPM	0.912	0.925	0.931	0.924	0.913	0.909

Table 6 CPM score with different K on target domain TianChi

K	2	3	4	5	6	7
CPM	0.916	0.931	0.933	0.935	0.935	0.935

Table 7 CPM score with different T on target domain TianChi

T	0.01	0.05	0.10	0.20	0.50
CPM	0.922	0.928	0.931	0.918	0.916

detects the scale-invariant but consistent feature information, thus reducing the false positive rate (sensitivity is much higher at low FPs/scan rate); and (3) the joint use of USSS and ZZC in a semi-supervised method achieves the best CPM score, validating its model generalization ability when transferred from the source domain to the target domain.

5 Discussion

5.1 Hyper-parameter setting

The τ , K and T are three hyper-parameters. We conduct experiments to discuss how to choose them. Tables 5–7 show the results with different values of τ , K and T , respectively. For τ , which is used to transform the predicted soft pseudo mask into binary pseudo mask in USSS, all target domain samples tend to have high uncertainty when τ is too high, while conversely, smaller τ makes the sample uncertainty lower. Finally, we set τ to 0.35. Beyond that, K is the iteration number in USSS. We can find that it is able to stably filter out uncertain target domain samples when K is greater than 3. T is a temperature hyper-parameter in ZZC, which makes contribution to entropy regularization. In general, a larger T will have little effect on entropy minimization, while a smaller T will sharpen the pseudo mask, but probably introduce some noise. Based on the results, we can observe that our model is robust to T , and finally T is set to 0.10.

5.2 Swapping: TianChi as source domain and LUNA16 as target domain

To further validate the robustness of our proposed human-in-the-loop method, we swap the originally defined source domain and target domain, and select the TianChi dataset as source domain and select the LUNA16 dataset as target domain. Experimental results are shown in Table 8. From the results, we can see that when the model trained on the source domain TianChi dataset is directly applied to the target domain LUNA16 dataset, the CPM score drops significantly, which is mainly caused by the domain gap. The 7.4% performance gain between baseline (target) and baseline (source) indicates the importance of introducing target domain supervision. Moreover, although the UDA methods attempt to bridge the domain gap through image-to-image translation or agent task training, their performance is again far behind the SSL methods. One important reason is that the SSL methods use a small amount of but effective target domain supervision produced by our USSS. Finally, our method outperforms the other SSL methods with 1.1%~2.3% improvement, further demonstrating the effectiveness of our elaborately designed ZZC strategy.

6 Conclusion

It is a common phenomenon that a well-performed model may suffer from performance degradation due to the domain gap. Although UDA methods are designed to reduce domain variance, their performance is still limited because no target domain information is introduced. Instead, delineating a few target domain samples would be a direct and feasible way to bridge the gap. Following this idea, we propose a human-in-the-loop method for pulmonary nodule detection in a semi-supervised manner to boost the model's generalization ability when transferred from the source domain to the target domain. Specifically, we first train a well-performing model on the source domain, and then a few samples containing rich information from the target domain using our proposed USSS. Finally, both labeled and unlabeled target domain samples are utilized by ZZC to learn scale-invariant but consistent target domain features. Our method is verified on two pulmonary nodule datasets. Extensive experimental results show that our method can not only achieve the best performance in the source domain compared with other advanced networks, but also can achieve surprising results when transferred to the target domain, compared to competitive UDA and SSL methods.

However, the proposed HILT method needs to first discover some target domain samples worth annotating and then refine the pre-trained source domain model in an SSL fashion, which is a little complicated. Therefore, in our future work, we plan to develop a one-stage pulmonary nodule detection system using a more generalized HITL

Table 8 Performance comparison with the UDA and SSL methods on the target domain LUNA16 dataset. “UDA” and “SSL” are unsupervised domain adaptation and semi-supervised learning, respectively. Note that all SSL methods using the same 5% labeled target domain samples, which are selected by our proposed USSS

Method	Type	FPs/scan							CPM
		0.125	0.25	0.5	1	2	4	8	
Baseline (source)	–	0.632	0.701	0.752	0.813	0.846	0.853	0.860	0.780
Baseline (target)	–	0.677	0.786	0.835	0.886	0.916	0.935	0.943	0.854
UNet-GAN [30]	UDA	0.742	0.826	0.883	0.909	0.918	0.923	0.924	0.875
SFDA [23]	UDA	0.747	0.853	0.890	0.914	0.921	0.926	0.928	0.883
ConstraintAdap [22]	UDA	0.764	0.860	0.900	0.915	0.923	0.929	0.931	0.889
UA-MT [41]	SSL	0.802	0.869	0.893	0.918	0.927	0.932	0.935	0.896
CPS [34]	SSL	0.827	0.873	0.906	0.925	0.928	0.934	0.936	0.904
ST++ [58]	SSL	0.833	0.877	0.912	0.928	0.931	0.935	0.938	0.908
Ours	SSL	0.858	0.910	0.917	0.925	0.936	0.942	0.947	0.919

method, i.e., performing expert annotation and cross-domain detection simultaneously, with the goal of achieving perfect cross-domain pulmonary nodule detection.

Abbreviations

CAD, computer-aided diagnosis; CNNs, convolutional neural networks; CPM, competition performance metric; CT, computed tomograph; FC, feature-level constraint; FP/scan, false positives per scan; FROC, free-response receiver operating characteristic; GAN, generative adversarial network; GN, group normalization; HITL, human-in-the-loop; LUNA16, lung nodule analysis 16; MC, mutual consistency; MRI, magnetic resonance image; SSL, semi-supervised learning; UDA, unsupervised domain adaptation; USSS, uncertainty-guided sample selection scheme; ZZC, zoom-in and zoom-out constraint.

Author contributions

QZ and YX proposed the main idea of human-in-the-loop method for pulmonary nodule detection. QZ and ZL designed the training pipeline to validate the model performance. QZ, YX and ZL designed ablations to demonstrate the contribution of each model component. YX was the project leader, responsible for obtaining funding and participating in the entire process of methodology design and evaluation. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (No.62171377), in part by the National Key R&D Program of China (Nos.2022YFC2009903 and 2022YFC2009900), and in part by the Ningbo Clinical Research Center for Medical Imaging (No. 2021L003) (Open Project 2022LYKFZD06).

Availability of data and material

All data generated or analyzed during this study are included in this published article [27, 28].

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author details

¹National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China.
²Australian Institute for Machine Learning, The University of Adelaide, Adelaide SA 5000, Australia.

Received: 4 January 2024 Revised: 16 June 2024 Accepted: 17 June 2024
 Published online: 10 July 2024

References

- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249.
- Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71(1), 7–33.
- Dou, Qi, Chen, H., Jin, Y., Lin, H., Qin, J., & Heng, P.-A. (2017). Automated pulmonary nodule detection via 3D ConvNets with online sample filtering and hybrid-loss residual learning. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 630–638). Cham: Springer.
- Dou, Q., Chen, H., Yu, L., Qin, J., & Heng, P.-A. (2016). Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7), 1558–1567.
- Drokin, I., & Elicheva, E. (2021). End-to-end lung nodule detection framework with model-based feature projection block. In *Proceedings of the international workshop on machine learning in medical imaging* (pp. 91–100). Cham: Springer.
- Liu, W., Liu, X., Li, H., Li, M., Zhao, X., & Zhu, Z. (2021). Integrating lung parenchyma segmentation and nodule detection with deep multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 25(8), 3073–3081.
- Luo, X., Song, T., Wang, G., Chen, J., Chen, Y., Li, K., et al. (2022). SCPM-Net: an anchor-free 3D lung nodule detection network using sphere representation and center points matching. *Medical Image Analysis*, 75, 102287.
- Mei, J., Cheng, M.-M., Xu, G., Wan, L.-R., & Sanet, H. Z. (2021). A slice-aware network for pulmonary nodule detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8), 4374–4387.
- Song, T., Chen, J., Luo, X., Huang, Y., Liu, X., Huang, N., et al. (2020). CPM-Net: a 3D center-points matching network for pulmonary nodule detection in CT scans. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 550–559). Cham: Springer.
- Zeng, Q., & Geng, J. (2022). Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191, 143–154.
- Lu, Z., Lu, M., & Xia, Y. (2022). M2f: a multi-modal and multi-task fusion network for glioma diagnosis and prognosis. In *Proceedings of the international workshop on multiscale multimodal medical imaging* (pp. 1–10). Cham: Springer.
- Zhou, Z., Gou, F., Tan, Y., & Wu, J. (2022). A cascaded multi-stage framework for automatic detection and segmentation of pulmonary nodules in developing countries. *IEEE Journal of Biomedical and Health Informatics*, 26(11), 5619–5630.

13. Zeng, Q., Geng, J., Jiang, W., Huang, K., & Idin, Z. W. (2021). Iterative distribution learning network for few-shot remote sensing image scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.
14. Zheng, S., Guo, J., Cui, X., Veldhuis, R. N. J., Oudkerk, M., & Van Ooijen, P. M. A. (2019). Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. *IEEE Transactions on Medical Imaging*, 39(3), 797–805.
15. Guo, Z., Zhao, L., Yuan, J., & Msanet, H. Yu. (2021). Multiscale aggregation network integrating spatial and channel information for lung nodule detection. *IEEE Journal of Biomedical and Health Informatics*, 26(6), 2547–2558.
16. Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3), 1173–1185.
17. Wolterink, J. M., Dinkla, A. M., Savenije, M. H. F., Seevinck, P. R., van den Berg, C. A. T., & Išgum, I. (2017). Deep MR to CT synthesis using unpaired data. In *Proceedings of the international workshop on simulation and synthesis in medical imaging* (pp. 14–23). Cham: Springer.
18. Panfilov, E., Tiulpin, A., Klein, S., Nieminen, M. T., & Saarakkala, S. (2019). Improving robustness of deep learning based knee MRI segmentation: mixup and adversarial domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 450–459). Piscataway: IEEE.
19. Bermúdez-Chacón, R., Altingöyde, O., Becker, C., Salzmann, M., & Fua, P. (2019). Visual correspondences for unsupervised domain adaptation on electron microscopy images. *IEEE Transactions on Medical Imaging*, 39(4), 1256–1267.
20. Perone, C. S., Ballester, P., Barros, R. C., & Cohen-Adad, J. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194, 1–11.
21. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232). Piscataway: IEEE.
22. Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., & Ayed, I. B. (2021). Constrained domain adaptation for image segmentation. *IEEE Transactions on Medical Imaging*, 40(7), 1875–1887.
23. Liu, Y., Zhang, W., & Wang, J. (2021). Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1215–1224). Piscataway: IEEE.
24. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generations Computer Systems*, 135, 364–381.
25. Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 102062.
26. Liu, Z., Wang, J., Gong, S., Lu, H., & Tao, D. (2019). Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6122–6131). Piscataway: IEEE.
27. Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S. N., van den Bogaard, C., Cerello, P., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical Image Analysis*, 42, 1–13.
28. Tianchi dataset. Retrieved June 10, 2024, from: <https://tianchi.aliyun.com/competition/entrance/231601/information?from=oldUrl/>.
29. Zeng, Q., Geng, J., Huang, K., Jiang, W., & Guo, J. (2021). Prototype calibration with feature generation for few-shot remote sensing image scene classification. *Remote Sensing*, 13(14), 2728.
30. Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., & Tao, Q. (2019). The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 623–631). Cham: Springer.
31. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 234–241). Cham: Springer.
32. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., & Xia, Y. (2023). Segment together: a versatile paradigm for semi-supervised medical image segmentation. arXiv preprint. [arXiv:2311.11686](https://arxiv.org/abs/2311.11686).
33. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., & Ding, X. (2020). Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63, 101693.
34. Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2613–2622). Piscataway: IEEE.
35. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., et al. (2022). Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81, 102530.
36. Zeng, Q., Xie, Y., Lu, Z., & Pefat, Y. X. (2023). Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15671–15680). Piscataway: IEEE.
37. Sajjadi, M., Javanmardi, M., & Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 1163–1171). Red Hook: Curran Associates.
38. Xie, Y., Zhang, J., Liao, Z., Verjans, J., Shen, C., & Xia, Y. (2022). Intra- and inter-pair consistency for semi-supervised gland segmentation. *IEEE Transactions on Image Processing*, 31, 894–905.
39. Zeng, Q., Xie, Y., Lu, Z., Lu, M., & Xia, Y. (2023). Discrepancy matters: learning from inconsistent decoder features for consistent semi-supervised medical image segmentation. arXiv preprint. [arXiv:2309.14819](https://arxiv.org/abs/2309.14819).
40. Miyato, T., Maeda, S., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8), 1979–1993.
41. Yu, L., Wang, S., Li, X., Fu, C.-W., & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 605–613). Cham: Springer.
42. Luo, X., Chen, J., Song, T., & Wang, G. (2021). Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8801–8809). Palo Alto: AAAI Press.
43. Wu, Y., Xu, M., Ge, Z., Cai, J., & Zhang, L. (2021). Semi-supervised left atrium segmentation with mutual consistency training. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 297–306). Cham: Springer.
44. Taleb, A., Lippert, C., Klein, T., & Nabi, M. (2021). Multimodal self-supervised learning for medical image analysis. In *Proceedings of the 27th international conference on information processing in medical imaging* (pp. 661–673). Cham: Springer.
45. Wang, H., Chen, T., Wang, Z., & Ma, K. (2023). Troubleshooting image segmentation models with human-in-the-loop. *Machine Learning*, 112(3), 1033–1051.
46. Ravanbakhsh, M., Tschernetzki, V., Last, F., Klein, T., Batmanghelich, K., Tresp, V., et al. (2020). Human-machine collaboration for medical image segmentation. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 1040–1044). Piscataway: IEEE.
47. Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint. [arXiv:1411.1784](https://arxiv.org/abs/1411.1784).
48. Quan, Q., Yao, Q., Li, J., & Zhou, S. K. (2022). Which images to label for few-shot medical landmark detection? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20606–20616). Piscataway: IEEE.
49. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway: IEEE.
50. Wu, Y., & He, K. (2018). Group normalization. In Ferrari, V., Hebert, M., Sminchisescu, C., et al. (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 3–19). Cham: Springer.
51. Venturini, L., Papageorgiou, A. T., Noble, J. A., & Namburete, A. I. L. (2020). Uncertainty estimates as data selection criteria to boost omni-supervised learning. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 689–698). Cham: Springer.
52. Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), *Proceedings of the 34th international conference on*

- neural information processing systems (pp. 6256–6268). Red Hook: Curran Associates.
53. Cao, H., Liu, H., Song, E., Ma, G., Xu, X., Jin, R., et al. (2020). A two-stage convolutional neural networks for lung nodule detection. *IEEE Journal of Biomedical and Health Informatics*, 24(7), 2006–2015.
 54. Bandos, A. I., Rockette, H. E., Song, T., & Gur, D. (2009). Area under the free-response roc curve (froc) and a related summary index. *Biometrics*, 65(1), 247–256.
 55. Niemeijer, M., Loog, M., Abramoff, M. D., Viergever, M. A., Prokop, M., & van Ginneken, B. (2010). On combining computer-aided detection systems. *IEEE Transactions on Medical Imaging*, 30(2), 215–223.
 56. Wang, D., Zhang, Y., Zhang, K., & Focalmix, L. W. (2020). Semi-supervised learning for 3D medical image detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3951–3960). Piscataway: IEEE.
 57. Gao, S., Li, Z.-Y., Yang, M.-H., Cheng, M.-M., Han, J., & Torr, P. (2023). Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), 7457–7476.
 58. Yang, L., Zhuo, W., Qi, L., Shi, Y., & Gao, Y. (2022). St++: make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4268–4277). Piscataway: IEEE.
 59. van Ginneken, B., Armato, S. G. III., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., et al. (2010). Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study. *Medical Image Analysis*, 14(6), 707–722.
 60. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In I. Guyon, U. von Luxburg, S. Bengio, et al. (Eds.), *Proceedings of the 31st international conference on neural information processing systems* (pp. 5574–5584). Red Hook: Curran Associates.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)