Applied Intelligence
https://doi.org/10.1007/s10489-021-03025-7

# Semantic consistency generative adversarial network for cross-modality domain adaptation in ultrasound thyroid nodule classification

**Jun Zhao[1] · Xiaosong Zhou[1] · Guohua Shi[1] · Ning Xiao[1] · Kai Song[1] · Juanjuan Zhao[1] · Rui Hao[2] · Keqin Li[3]**

Accepted: 18 November 2021


**Abstract**
Deep convolutional networks have been widely used for various medical image processing tasks. However, the performance of existing learning-based networks is still limited due to the lack of large training datasets. When a general deep model is directly deployed to a new dataset with heterogeneous features, the effect of domain shifts is usually ignored, and performance degradation problems occur. In this work, by designing the semantic consistency generative adversarial network (SCGAN), we propose a new multimodal domain adaptation method for medical image diagnosis. SCGAN performs cross-domain collaborative alignment of ultrasound images and domain knowledge. Specifically, we utilize a self-attention mechanism for adversarial learning between dual domains to overcome visual differences across modal data and preserve the domain invariance of the extracted semantic features. In particular, we embed nested metric learning in the semantic information space, thus enhancing the semantic consistency of cross-modal features. Furthermore, the adversarial learning of our network is guided by a discrepancy loss for encouraging the learning of semantic-level content and a regularization term for enhancing network generalization. We evaluate our method on a thyroid ultrasound image dataset for benign and malignant diagnosis of nodules. The experimental results of a comprehensive study show that the accuracy of the SCGAN method for the classification of thyroid nodules reaches 94.30%, and the AUC reaches 97.02%. These results are significantly better than the state-of-the-art methods.

**Keywords** Cross-modality domain adaptation · Semantic consistency · Domain knowledge · Self-attention mechanism · Thyroid nodule classification

## 1 Introduction

Thyroid nodules, described as abnormal growths of glandular tissue, are the most common thyroid disorder [2]. Over the past 30 years, thyroid cancer has been one of the most prevalent and fastest-growing cancers of all types [15]. Therefore, early diagnosis of the benignity or malignancy of nodules is essential to reduce the morbidity and mortality of thyroid cancer [8]. Ultrasonography has become the most preferred choice for diagnosing benign and malignant thyroid nodules. However, there are still some challenges in analyzing thyroid ultrasound images. First, ultrasound images are susceptible to speckle noise and echo fluctuations, making the texture distribution in ultrasound images blurred and non-uniform [37]. Second, the diagnosis of thyroid ultrasound images is subjective and highly dependent on the physicians' extensive experience and cognitive ability [14]. Conversely, the use of computer-aided diagnosis systems (CADs) can significantly reduce physicians' workload and misdiagnosis rate. Thyroid image classification has become a research hotspot for computer-aided thyroid disease diagnosis [5].

**Traditional methods of thyroid nodule classification** Acharya et al. [1] used Gabor transform to extract the features of thyroid benign and malignant images and compared

✉ Juanjuan Zhao
zhaojuanjuan@tyut.edu.cn

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, China

[2] College of Information, Shanxi University of Finance and Economics, Taiyuan, China

[3] Hunan University and State University of New York, Albany, NY, USA

Published online: 13 January 2022

the classification performance of SVM, MLP, KNN, and C4.5 classifiers. Raghavendra et al. [26] extracted high-order spectral (HOS) entropy features from particle swarm optimization (PSO) and support vector machine (SVM) models and distinguished benign and malignant lesions. Prochazka et al. [23] used dual-threshold binary decomposition to extract direction-independent features for random forest (RF) and SVM classifiers. The traditional training method is computationally inexpensive and does not require a large number of training images. However, there are still apparent limitations: 1) rely on many manually extracted image features and classifier selection, 2) is a tedious and unstable process, and 3) may lead to poor generalization ability.

**Thyroid nodules classification method based on deep learning** Compared with traditional methods, deep learning methods can extract global and local features more accurately. In 2017, Ma et al. [16] applied the convolutional neural network for the first time to identify benign and malignant thyroid nodules. Wang et al. [36] designed an effective EM algorithm to train a CNN-based nodule classification model. Zhou et al. [50] proposed an online transfer learning (OTL) method to improve the diagnostic effect of ultrasound examination of thyroid nodules. Wang et al. [37] extracted multiple image features with different angles in one inspection for an attention-based feature aggregation network.

All the above methods are based on single modality data for training and evaluation. In contrast, the actual medical imaging process expects to fuse data from different domains. Still, the following problems exist in the construction of models: 1) The scale of medical datasets remains a significant bottleneck for deep learning models. Data collection and manual annotation for each new modality or new domain are both time-consuming and expensive. Especially for thyroid imaging, there are fewer extant large-scale thyroid image datasets due to the specificity of thyroid location. 2) The distribution differences between different types of data, known as dataset deviations or domain shifts phenomenon, where deep networks trained on a large labeled dataset cannot be well generalized to new datasets and new tasks, resulting in significant degradation of the generalization performance the model.

We adopt a domain adaptation (DA) algorithm [38] to address the above challenges. The DA algorithm aims to learn models from the source domain data distribution but works well for target domains with different but related data distribution. The principle behind DA is that the source and target domains can learn collaboratively and transfer their learned knowledge to each other during the entire training process, making the model robust to noise in the data. Currently, there is no work on effectively using cross-modal data to construct a DA framework for nodule diagnosis in thyroid ultrasound images.

In general, the working pattern of the ultrasound physician is to combine information from both ultrasonography reports and ultrasound images and then to come up with a diagnosis. This model stimulated our interest in exploring the content of the reports. We find that the performance of image generation and image classification tasks can be improved by transferring the semantic-intensive feature representation associated with the images in the reports. In contrast, existing models lack the reasoning ability to imitate a physicians' interpretation of semantic information and ignore important domain and expert knowledge [41] related to the specific task of thyroid diagnosis. Therefore, in our approach, we will incorporate disease keywords extracted from ultrasonography reports as textual information in multimodal data, as shown in Fig. 1.

In this paper, we propose a new multi-task cascaded deep learning framework for diagnosing thyroid ultrasound images. First, we propose a self-attention-based semantic consistency generative adversarial network as a domain adaptation backbone to improve the quality of generated images. Second, to jointly analyze multimodal data features, the critical domain knowledge extracted from ultrasonography reports is fed into the generator structure through text modeling to promote the semantic consistency of generated images. Finally, the network integrates a modified classification model, ResNet-50, which uses combined features to classify benign or malignant thyroid nodules in ultrasound images.

The main contributions of this paper are summarized as follows:

- We propose an effective model: semantic consistency generative adversarial network (SCGAN). To the best of our knowledge, this work is the first to apply cross-modal domain adaptation based on generative adversarial networks to the classification task of benign and malignant thyroid nodules.
- We propose a new cross-modal alignment self-attention module (CASAM) to facilitate domain adaption for achieving higher generative performance. The semantic alignment layer is used in CASAM to efficiently guide the semantic alignment process of image and knowledge features.
- We introduce two advanced techniques: the visual discrepancy loss to dynamically balance the need for the generator to learn domain invariant features, and the cross-domain fusion zero-centered gradient penalty (CD-GP) is incorporated into the discriminator to synthesize more realistic and knowledge semantic consistent images.

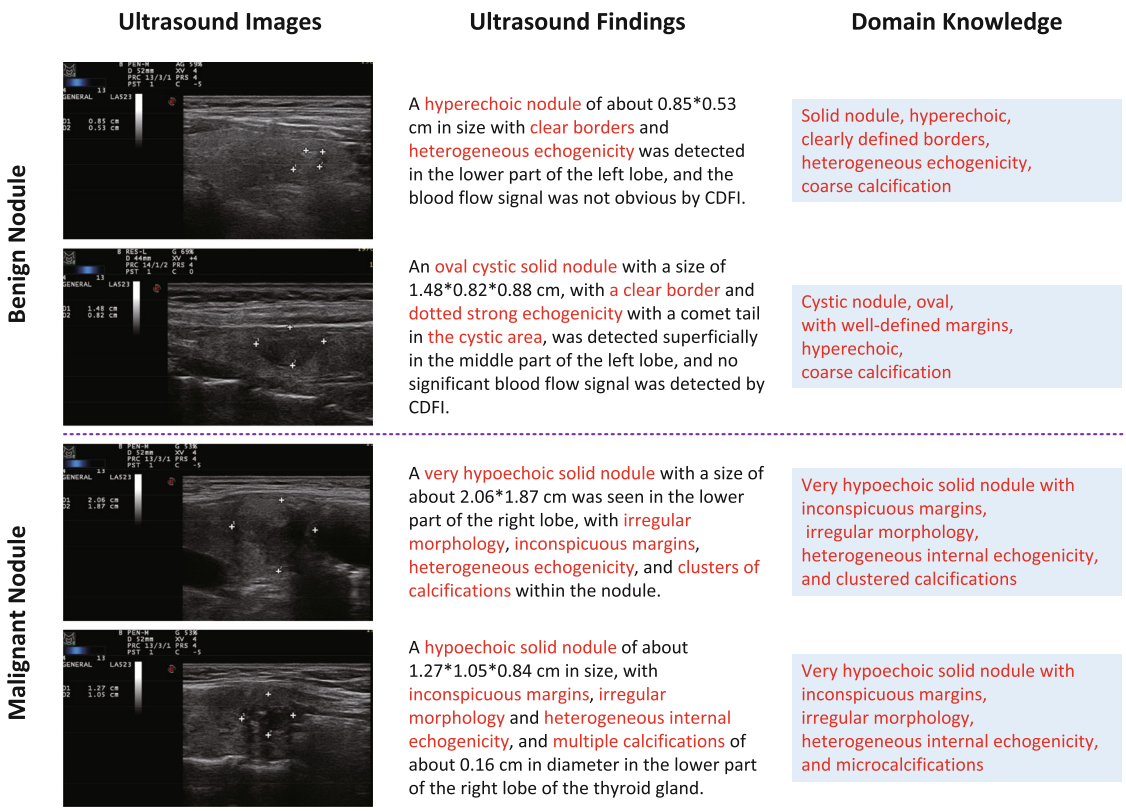


**Fig. 1** From left to right, the original ultrasound images of the four benign/malignant nodules randomly sampled in the dataset, the corresponding "Ultrasound Findings" in the ultrasonography report, and the domain knowledge are shown. Among them, the red text description is based on the relevant disease keywords selected by TI-RADS as the standard

- Extensive experiments show that our proposed SCGAN achieves good results in thyroid nodules' ultrasound image generation task and is well validated in the image classification results.

The rest of the paper is organized as follows. We present related work on domain adaption, generative adversarial networks, and attention mechanisms for medical images in Section 2. The details of our approach are presented in Section 3. Section 4 describes our thyroid ultrasound image dataset and experimental evaluation results to validate the effectiveness of our approach. Finally, the conclusion and future work are drawn in Section 5.

## 2 Related work

**Domain Adaptation** In the context of medical image analysis, most prospective studies on domain adaptation have focused on adjusting data distribution from various clinical centers, scanning protocols, and scanning sites. Dou et al. [6] pioneered a plug-and-play adversarial domain adaptation network (PnP-AdaNet), which combines multiple adversarial learning domain adaptation layers to spatially align the potential features of the target domain and the source domain. They tested on cardiac MRI/CT images. Zhang et al. [49] introduced a collaborative unsupervised domain adaptation (CoUDA) algorithm for medical image diagnosis. This algorithm via the collective intelligence of two peer subnets to conduct transferability-aware domain adaptation on whole-slide images (WSI) and microscopy images (MSI) of colon datasets. However, it is often difficult to seek a source domain with the same feature and categorical space as the target domain. Therefore, this paper focuses on more realistic and challenging scenarios to address the correlation problems of cross-domain data observed in different feature spaces, namely heterogeneous domain adaptation (HDA) [44].

**Generative adversarial network** The domain-invariant representation of classification tasks from the source dataset to the target dataset has been extensively studied [38] by generating adversarial networks [45]. Chen et al. [7] investigated the domain adaptation framework, SIFA, which applies a deep supervision mechanism of synergistic image and feature alignment to deal with the transfer of domains due to adversarial learning, and extensive experiments on bidirectional cross-modality adaptation on multiple tasks. Ren

et al. [27] considered the joint feature distribution between the source and target domain images and classified histological images obtained in different staining procedures via adversarial learning. Gu et al. [11] explored a two-step progressive transfer learning technique to improve the recognition performance of cross-domain skin diseases, and at the same time, adopted cycle-consistent adversarial learning to expand the model to cross-modal learning tasks such as melanoma detection.

**Attention Mechanism** Although existing adversarial domain adaptation methods are effective in different tasks, the semantic correlation between domains has not been elucidated yet. Nowadays, attention mechanism has become a necessary element to capture inter-domain dependencies of the model. Wang et al. [40] added transferable attention for the domain adaptation (TADA) model and focused its application on core regions to enhance the transferability of images. Wang et al. [34] argued that complementing the attention branch in the Thorax-Net enhances the correlation between class labels and pathological abnormal locations. Furthermore, the three attention modules [35] can be merged into a unified framework for joint learning of channels, elements, and scales. In the thyroid ultrasound nodule diagnosis, we will demonstrate an improved version of a well-established self-attention mechanism to improve further diagnostic performance, which helps localize important regions of ultrasound images and enhancing cross-domain features' correlatability.

## 3 Methods

This section illustrates the proposed semantic consistency generative adversarial network (SCGAN) for ultrasound image nodule classification. First, we introduce the selection criteria of domain knowledge and the processing of its integration into deep networks. Second, we present the overall structure of SCGAN, including the composition of the generator and discriminator, and focus on the contribution of the cross-modal alignment self-attention module to semantic consistency. Then, we explain the proposed visual discrepancy loss and regularization method. Finally, we give details of the modifications of the classifier.

### 3.1 Domain knowledge

**Ultrasonography report preprocessing** The ultrasonography report [13] summarizes all clinical findings and physician impressions identified during the ultrasound study examination. Ultrasonography reports usually contain comprehensive patient information, but they may also contain inconclusive descriptions or irrelevant to the disease. For example, in the "Ultrasound Findings" of the ultrasonography report, as shown in Fig. 1, normal/abnormal conditions are recorded for each site of the thyroid examination, such as location, size, and severity of the nodules. Besides, patients' personal information, medical history, and suspicious findings may lead to additional or follow-up studies. Therefore, parsing the content of ultrasonography reports is a complex and challenging task.

The Thyroid Imaging Reporting and Data System (TI-RADS) [32] provides standardized terminology to describe thyroid nodule features in ultrasound images. Using TI-RADS as a guide, we screen the disease keywords in ultrasonography reports as domain knowledge, such as boundary, calcification, and echo pattern. By learning text embedding, this domain knowledge can facilitate the acquisition of semantic information in ultrasonography reports and improve the diagnostic performance of the leading classification tasks.

**LSTM for Text Modeling** We use a pre-trained text encoder $\phi$ to learn the semantic information described by domain knowledge. Each textual description $t_i$ is encoded as one-hot vectors that are then mapped to embeddings and added with contextual information. The text embedding $\phi(t_i)$ is fed into the LSTM proposed by [10]. At each time step, the obtained text embedding sequence $\{\phi_1, ..., \phi_n\}$ takes the current text as input and iteratively applies the transfer function to generate the hidden state $h_t$:

$$h_t = LSTM(\phi_t, h_{t-1}) \tag{1}$$

which allows the extraction of high-dimensional semantic vectors from domain knowledge. Domain knowledge contains meaningful disease features, and the key is to maintain diversity and independence among them. To this end, we extract the hidden state corresponding to each disease keyword and obtain a text representation sequence $T_s^{enc} = [h_1, ..., h_n] \in R^{\mathcal{E}}$. The advantage of this strategy is that it enables the network to select relevant semantic features adaptively, ensuring that they are helpful for disease labeling (as shown in the experimental results).

### 3.2 Semantic consistency generation adversarial network

#### 3.2.1 A. Model overview

Our network architecture is shown in Fig. 2. It consists of a pre-trained text encoder, a domain adaptation generator, and a discriminator. The generator is trained to generate images from the text describing the content, and the discriminator is trained to determine the authenticity of the images conditional on the semantics defined by the given text. We use the following notations: the domain adaptation

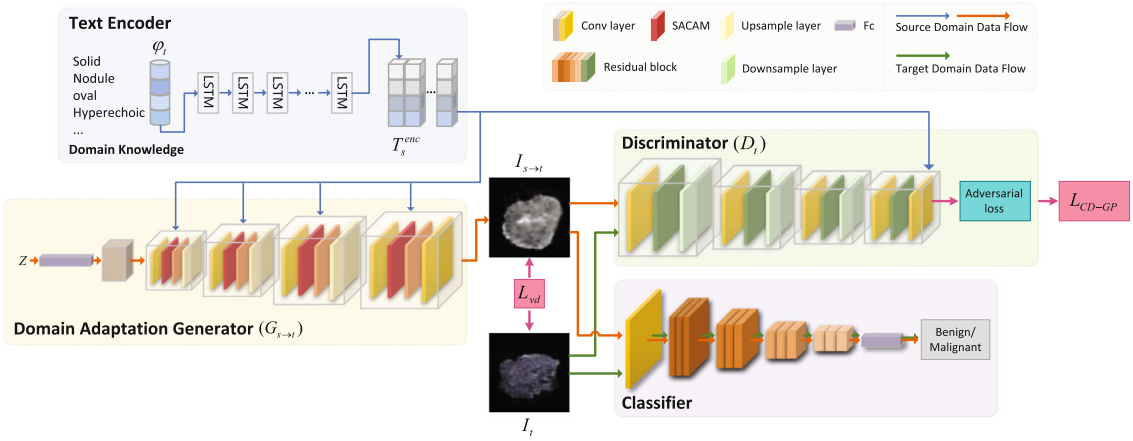


**Fig. 2** Overview of our proposed SCGAN, consisting of a text encoder (top left), a domain adaptation generator $G_{s\to t}$(bottom left), a discriminator $D_t$(top right), and a classifier (bottom right). $G_{s\to t}$ has two inputs, $Z$ and $T_s^{enc}$ generated by the text encoder, both of which are implemented in upblocks (gray boxes) for cross-domain fusion. The SACAM contained in upblocks promotes semantic alignment during the fusion process. Similarly, $D_t$ distinguishes the authenticity of an image by a series of downblocks (gray boxes, $I_{s\to t}$ represents the synthesized image, $I_t$ represents the real image). The classifier is the modified classification model ResNet-50. In particular, the adversarial loss refers to the hinge version of the adversarial loss, $\mathcal{L}_{vd}$ is the visual discrepancy loss, and $\mathcal{L}_{CD-GP}$ is the cross-domain fusion zero-centered gradient penalty function

generator is denoted as $G_{s\to t}$: $R^{\mathcal{Z}} \times R^{\mathcal{E}} \to R^{\mathcal{D}}$, the discriminator is denoted as $D_t$: $R^{\mathcal{D}} \times R^{\mathcal{E}} \to \{0,1\}$, where $\mathcal{E}$ is the dimension of the embedded text representation, $\mathcal{D}$ is the dimension of the image, and $\mathcal{Z}$ is the dimension of the noise input in $G_{s\to t}$.

The generator $G_{s\to t}$ has two inputs, the text sequence $T_s^{enc}$ of the source domain, and the other is the noise vector $Z \in R^{\mathcal{Z}} \sim \mathcal{N}(0, 1)$ sampled from the Gaussian distribution to guarantee the diversity of the generated images. First, $Z$ is fed into the fully connected layer and then sent to a series of upblocks and $T_s^{enc}$ to upsample the images, which are used to integrate semantic information and image features during the image generation process. $G_{s\to t}$ uses upblocks as its network backbone, including convolutional layers, a self-attention layer, residual blocks, and an upsample layer. The self-attention layer brings more non-linearity to $G_{s\to t}$, which is conducive to generating semantically consistent images from different textual descriptions. Therefore, $G_{s\to t}$ synthesizes realistic pseudo-target domain images by $I_{s\to t} = G_{s\to t}(Z, T_s^{enc})$. Then $I_{s\to t}$ is regularized using visual discrepancy loss to be consistent with the corresponding region in the original image.

The discriminator $D_t$ attempts to compete with $G_{s\to t}$ by distinguishing between the synthetic pseudo target domain image $I_{s\to t}$ and the real target domain image $I_t$. $D_t$ converts $I_{s\to t}$ into a feature map and downsamples it through a series of downblocks. Here, the intermediate layers of $D_t$ have a smaller receptive field that forces $G_{s\to t}$ to pay more attention to finer details. The last few layers generally derive information from the larger image region and guide $G_{s\to t}$ to produce an image with better global consistency. Then $T_s^{enc}$ is replicated and spliced onto the image features. Formally, $D_t$ has to distinguish three input pairs composed of text: real images $I_t^{match}$ with matching text, real images $I_t^{mis}$ with mismatched text, and synthetic images $I_{s\to t}$.

### 3.2.2 B. Cross-modal alignment self-attention module

For the data heterogeneity between source-domain text representation and target-domain images, we propose the cross-modal alignment self-attention module (CASAM). The self-attention module efficiently computes long-range dependencies between features, allowing the generator to model the relationship between widely separated spatial regions effectively. CASAM leverages semantic association to effectively guide the alignment process while generating attention to important image features and text representation to provide more prominent and meaningful embedding for image generation tasks.

As shown in Fig. 3, the module accepts two inputs: image feature map $F_i$ and text representation sequence $T_s^{enc}$. First, according to the attention mechanism adopted in AttnGAN [42], the three-dimensional image features (width× height×channel) of $F_i \in R^{w\times h\times c}$ are flattened into a two-dimensional sequence (wh×channel, where wh=width×height), and transformed into the query feature map $Q_{s\to t}$ to facilitate the calculation of attention. The formula is as follows:

$$Q_{s\to t} = W_i \cdot F_i \tag{2}$$

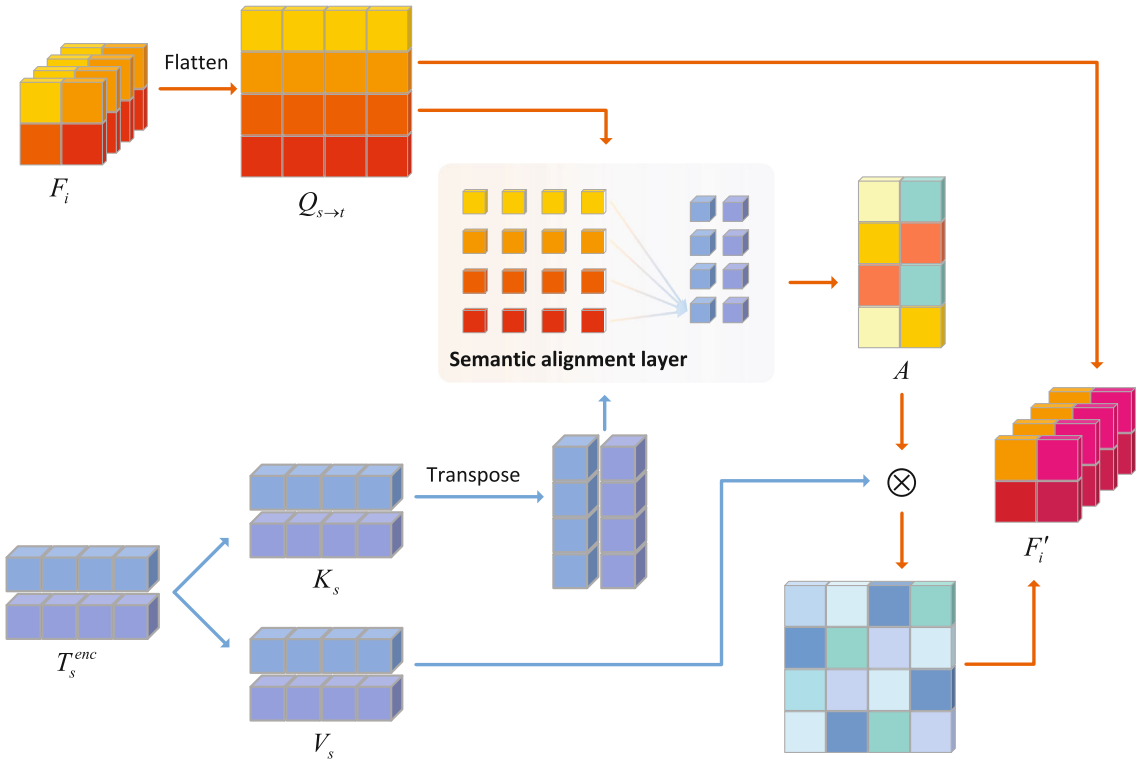


**Fig. 3** Details of the cross-modal alignment self-attention module (CASAM). The semantic alignment layer can focus on the source domain features corresponding to the target domain pixels. ⊗ denotes dot product operation

Two convolution layers with $1 \times 1$ filters are applied on $T_s^{enc}$ to generate feature maps $K_s$ and $V_s$, respectively:

$$K_s = W_k \cdot T_s^{enc} \tag{3}$$

$$V_s = W_v \cdot T_s^{enc} \tag{4}$$

Intuitively, the key $K_s$ focuses on matching with $Q_{s\to t}$, while the other projection value $V_s$ can be better optimized to refine $Q_{s\to t}$ to obtain better $F_i$.

We add a semantic alignment layer (SAL) to the module to strengthen the semantic relevance between $Q_{s\to t}$ and $K_s$ by metric learning [22]. Here, we use:

$$S = \arg\max \cos(Q_{s\to t}, K_s) \tag{5}$$

as the geometric similarity to measure the relationship between the potential feature space of $Q_{s\to t}$ and $K_s$. In consideration of building a reasonable distance metric [4, 19, 22], the cosine similarity [9, 18, 33] is chosen in this paper as:

$$\cos(Q_{s\to t}, K_s) = \frac{\sum_{i,j}(Q_{s\to t}(i,j), K_s(i,j))}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} Q_{s\to t}(i,j)^2}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} K_s(i,j)^2}} \tag{6}$$

The cosine similarity focuses on the similarity description of semantic classes. For the feature vector of each image subregion of $Q_{s\to t}$, the better the alignment, the shorter the distance.

The attention maps weight to the feature maps $Q_{s\to t}$ and $K_s$ are generated to achieve more discriminative feature representation, and the attention map $A$ is obtained as:

$$A = \frac{\exp(W_A \cdot S_k)}{\sum_{K=1}^{N}\exp(W_A \cdot S_k)} \tag{7}$$

The aggregation operation is defined as follows:

$$F_i' = \gamma(softmax(A \cdot V_s)) + F_i \tag{8}$$

where the more refined features are captured by the dot product between $A$ and $V_s$ for feature adaptation. The obtained attention weights are normalized using the softmax function to convert the values into relative probabilities. The features are updated by collecting the attention weights of each acquired feature and the original feature mapping to obtain contextual information.

### 3.2.3 C. Visual discrepancy loss

We propose a new visual discrepancy loss for the generator. Visual discrepancy loss is encouraged to capture disparity features. If there is no discrepancy loss, then the requirement for $G_{s\to t}$ to learn the invariant domain information will be weaker. Thus, co-training visual discrepancy loss is an implicit facilitator for improving network adaptation and plays a crucial role in improving the quality and consistency

of the final generated images. The $L2$ norm of the feature mapping between the real image $I_t$ and the generated image $I_{s\to t}$ is defined as:

$$\mathcal{L}_{vd} = \sum_{j=1}^{J}\{E_{I_{s\to t}\sim\mathbb{P}_g, I_t\sim\mathbb{P}_r}[\parallel \phi_j(I_{s\to t}) - \phi_j(I_t) \parallel_2]\} \quad (9)$$

where $\phi_j(\cdot)$ represents the process of extracting image feature maps.

#### 3.2.4 D. Cross-domain fusion zero-centered gradient penalty

Recently, Mescheder et al. [17] introduced a zero-centered gradient penalty, adding regular terms to make the discriminator apply zero-centered gradient penalty to the input. Extending it to our domain adaptation task. We propose a cross-domain fusion zero-centered gradient penalty (CD-GP) function to improve the discriminator's generalization capability. We choose to impose penalty terms on the real and generated data, respectively:

$$\begin{aligned}\mathcal{L}_{CD-GP} &= \alpha E_{I_t\sim\mathbb{P}_r}[\parallel \nabla_I D_t(I_t, T_s^{enc}) \parallel_2 + \parallel \nabla_T D_t(I_t, T_s^{enc}) \parallel_2] \\ &\quad +\beta E_{I_{s\to t}\sim\mathbb{P}_g}[\parallel \nabla_I D_t(I_{s\to t}, T_s^{enc}) \parallel_2 + \parallel \nabla_T D_t(I_{s\to t}, T_s^{enc}) \parallel_2] \quad (10)\end{aligned}$$

where $\alpha$ and $\beta$ are hyperparameters that balance the effectiveness of the gradient penalty and cannot both be zero.

Compared with adding discriminators to ensure the semantic consistency of the generated images, our CD-GP does not introduce additional networks to compute the semantic similarity and therefore does not increase the complexity of the domain adaptation process or the training parameters.

#### 3.2.5 E. Objective function

To stabilize and converge the training process of SCGAN, inspired by the SAGAN architecture [47], we evaluate the authenticity of the generated images and their consistency with the input semantics by minimizing the hinge version of the adversarial loss [3]. Formally, we represent the two outputs of $D_t$ as: $D_t^u(\cdot)$, the unconditional image score, and $D_t^c(\cdot)$, the conditional image score. Correspondingly, the objective functions $\mathcal{L}^D$ for $D_t$ are formulated as $\mathcal{L}^D_{uncond}$ and $\mathcal{L}^D_{cond}$, respectively:

$$\mathcal{L}^D_{uncond} = -E_{I_t\sim\mathbb{P}_r}[\log(D_t^u(I_t))] - E_{I_{s\to t}\sim\mathbb{P}_g}[\log(1 - D_t^u(I_{s\to t}))] \quad (11)$$

$$\begin{aligned}\mathcal{L}^D_{cond} = \; & -E_{I_t\sim\mathbb{P}_r}[\min(0, -1 + D_t^c(I_t^{match}, T_s^{enc}))] \\ & -E_{I_t\sim\mathbb{P}_{mis}}[\min(0, -1 - D_t^c(I_t^{mis}, T_s^{enc}))] \\ & -E_{I_{s\to t}\sim\mathbb{P}_g}[\min(0, -1 - D_t^c(I_{s\to t}, T_s^{enc}))]\end{aligned} \quad (12)$$

$\mathbb{P}_r$ is the real data distribution, $\mathbb{P}_g$ is the generated data distribution, and $\mathbb{P}_{mis}$ is the mismatching data distribution.

On the other side, $G_{s\to t}$ is trained to generate images that could trick $D_t$ into giving high scores on visually realistic images and match the text. Similarly, the objective functions $\mathcal{L}^G$ to be minimized by $G_{s\to t}$ are $\mathcal{L}^G_{uncond}$ and $\mathcal{L}^G_{cond}$, respectively:

$$\mathcal{L}^G_{uncond} = -E_{I_{s\to t}\sim\mathbb{P}_g}[\log(D_t^u(I_{s\to t}))] \quad (13)$$

$$\mathcal{L}^G_{cond} = -E_{I_{s\to t}\sim\mathbb{P}_g}[D_t^c(I_{s\to t}, T_s^{enc})] \quad (14)$$

Taking into account the adversarial loss, visual discrepancy loss, and cross-domain fusion zero-centered gradient penalty, our total loss is defined as the weighted sum of these losses, as follows:

$$\mathcal{L}_{Total} = \mathcal{L}^G + \lambda_1\mathcal{L}_{vd} + \mathcal{L}^D + \lambda_2\mathcal{L}_{CD-GP} \quad (15)$$

$\lambda_1$ and $\lambda_2$ are regularization parameters to balance the trade-off between $\mathcal{L}_{vd}$, $\mathcal{L}_{CD-GP}$, and other terms.

### 3.3 Modified classification model ResNet-50

Each residual block of the ResNet-50 [12] network uses a bottleneck structure, which helps overcome the problem of gradient disappearance in large models. To adapt the ResNet-50 network to our problem of classifying benign and malignant nodules, the base layer of the model is frozen, and then custom layers are added to form the final framework. Therefore, we remove its last fully connected layer and add three fully connected layers of 2048, 1024, and 2 neurons, respectively. The weights of the final fully connected layer are fine-tuned by using a back-propagation technique which uses a gradient descent optimization algorithm to minimize the cost function. The final output of the model is obtained using the sigmoid activation function.

## 4 Result

### 4.1 Datasets

Our research works use images from a dataset provided by the local hospital to acquire ultrasound examination images and ultrasonography reports of 1083 patients, and the hospital institutional review board approves the entire collection process Due to the variable size of nodules, we exclude nodules with tumor size $< 0.60$ cm or $> 3.00$ cm and finally include 1937 nodules from ultrasound examinations in the final analysis. Their available ultrasonography reports correlate with the ultrasound findings of 867 patients. Ultrasound images are screened by experienced thyroid ultrasound physicians (physicians with more than eight years of experience in thyroid ultrasound imaging) based on suspicious features in TI-RADS, solid components, hypoechoic, or markedly hypoechoic, microgranular or irregular margins, microcalcifications, and ultra-wide shapes. "Ultrasound Findings" are classified into two categories: benign or malignant. There are 1032 benign nodules

**Table 1** Distribution of data in our dataset

| | Benign | Malignant | Total |
|---|---|---|---|
| Training Set | 865 | 754 | 1619 |
| Validation Set | 97 | 84 | 181 |
| Test Set | 70 | 67 | 137 |
| Total | 1032 | 905 | 1937 |

and 905 malignant nodules. We use a nested 10-fold cross-validation independent evaluation model. The dataset is divided into the training set, validation set, and test set. The training and test datasets are divided by patients, and there is no overlap between the two datasets. The training set and the validation set consist of 1800 images, and the training set isolates approximately 10% of the data as the validation set. The test set consists of 137 images. The data set distribution is shown in Table 1.

**Ultrasound Image Preprocessing** As shown in Fig. 4 to extract regions of interest (ROIs) containing nodules, the metadata text (e.g., information about the scanner, location, patient.) placed on the images are discarded to obtain the actual ultrasound image regions. We count the horizontal and vertical diameters of all nodules so that the nodule with the cross marker symbols is in the center of the patch image. It is finally decided to fill the patch size with zero to a square of 64 × 64 pixels size to maintain the image aspect ratio, and the pixels in the image are normalized to zero mean and unit variance.

## 4.2 Evaluation metrics

Classification results are quantitatively evaluated by the mean and standard deviation of the obtained accuracy, sensitivity/recall, specificity, and area under the receiver operating curve (AUC). In this paper, the inception score (IS) [31] is chosen to measure the quality of the images generated by SCGAN. IS is the classical metric for evaluating GAN. Since IS does not reflect whether the generated images depend well on the given text representation, we combine it with physician evaluation. The semantic consistency of SCGAN is evaluated by experienced ultrasound physicians comparing the generated images with the corresponding domain knowledge description. We consider that physicians need to perform two tasks: one is to discriminate the authenticity of the image and determine whether the image matches the corresponding semantic information; the other is to diagnose the benignity or malignancy of the nodule.

## 4.3 Implementation details

The entire network is implemented using the TensorFlow framework based on Python 3.6 and trained on a workstation with Ubuntu 18.04 LTS, 2.90 GHz Intel(R) Xeon(R) W-2102 CPU, and two NVIDIA GTX Titan XP GPUs. For the text encoder, the dimension $\mathcal{E}$ is set to 128, and the length of words is set to 30. In order to compare with previous work, the parameters of our text encoder are fixed during training. In the generator, the dimension of $\mathcal{Z}$ is set to 512. In the experiments, the network is trained using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. On our dataset, training is set up with 300 epochs and a minibatch size of 16. We choose

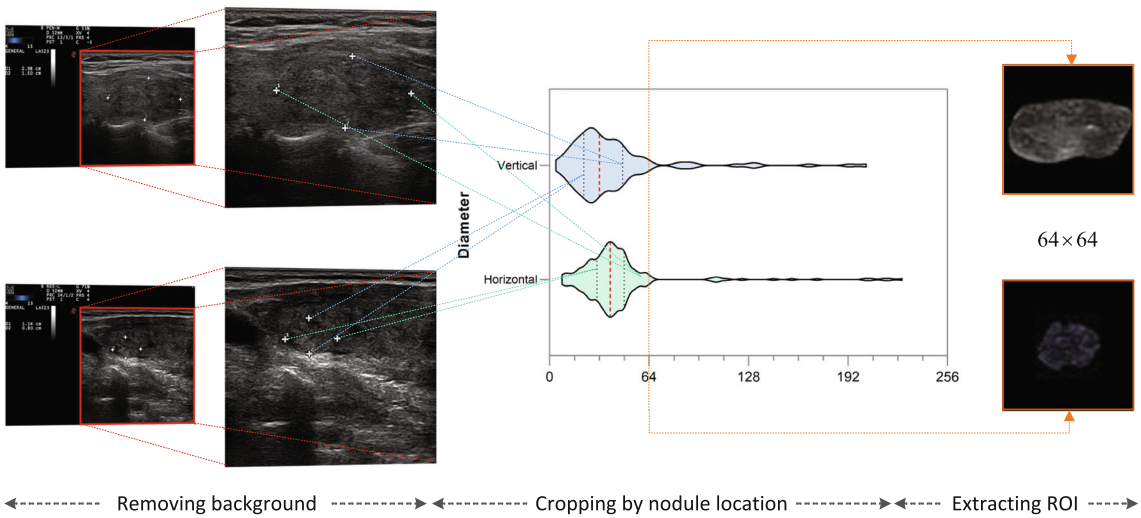


**Fig. 4** The overall process of image preprocessing. Where the blue dashed line indicates the vertical diameter of the nodule and the green dashed line indicates the horizontal diameter

a learning rate of $1e^{-3}$ for the classifier and $2e^{-4}$ for the rest of the architecture. The decay factor is 0.5 per 100 epochs. The target domain image enters the classifier for classification in the testing phase without involving GAN and other algorithm designs.

### 4.4 Experimental setup and analysis

For our proposed method, we set up three variants:

1. Remove the CASAM of SCGAN, additional loss functions $\mathcal{L}_{vd}$ and $\mathcal{L}_{CD-GP}$, that is, directly concatenating text representation with image features in $G_{s\to t}$ (DAGAN).
2. Use CASAM to fuse text representation and image features in $G_{s\to t}$ to test the contribution of CASAM in improving domain adaptation (SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$).
3. Only remove the visual discrepancy loss $\mathcal{L}_{vd}$ of SCGAN (SCGAN$-\mathcal{L}_{vd}$).

The effectiveness of our proposed method is demonstrated by designing several experimental sessions as follows.

In our research, the SCGAN model is based on an intelligent combination of knowledge and images. To evaluate the advantages of this cross-modal domain adaptation approach for feature extraction, first, we construct GANs model for nodule feature extraction using only images as input. The experimental results obtained by different classification methods are shown in Table 2 and Fig. 5. Here, the SCGAN model is simplified to DCGAN [25] when no domain knowledge is added, and only unimodal data is used for feature extraction. The accuracy, sensitivity, specificity, and AUC obtained using the DCGAN+modified ResNet-50 model are $85.26 \pm 1.62$, $87.46 \pm 3.14$, $83.14 \pm 1.69$, $84.80 \pm 2.51$, respectively. By adding class labels to DCGAN as auxiliary information to form ACGAN [20], the ACGAN+modified ResNet-50 model shows a slight improvement in all metrics. However, the classification performance of the above methods is far inferior to that of the GAN model using a multimodal combination of domain

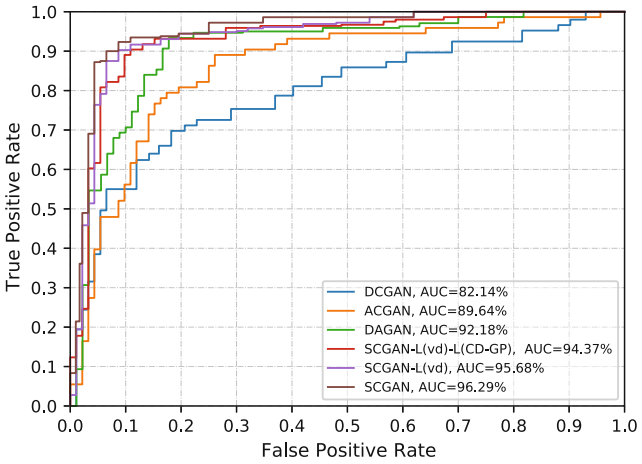


**Fig. 5** ROC analysis of different image generation models with our SCGAN and its variants for thyroid nodule classification

knowledge and images. Among them, the DAGAN model, the most basic variant of SCGAN, has metrics of $89.93 \pm 0.88$, $91.34 \pm 1.93$, $88.57 \pm 0.90$, $92.98 \pm 1.78$, respectively. Compared with DAGAN, the metrics of SCGAN are improved by another 4.37%, 2.09%, 6.59%, and 4.04%, respectively. It suggests that integrating the domain knowledge from ultrasonography reports into the deep learning model can effectively improve the classification performance of nodules. It can also be concluded that the standard deviation of the classification results is smaller when domain knowledge is used, which means that the inclusion of domain knowledge can effectively improve the stability of nodule classification. In addition, to verify the classification stability of SCGAN applied to unbalanced samples, we randomly reduce the number of malignant nodules by half, but the parameters of the fixed pre-training model remain unchanged, denoted as SCGAN$^\star$. In the case of unbalanced data sets, the fluctuation of each metric value is slight, and the classification performance of SCGAN is excellent.

**Ablation Study** To evaluate whether the self-attention mechanism can help the domain adaptation process to generate higher quality and semantically consistent images.

**Table 2** Comparison of the classification performance of different image generation models with our SCGAN

| Methods | Results(%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| DCGAN [25] | $85.26 \pm 1.62$ | $87.46 \pm 3.14$ | $83.14 \pm 1.69$ | $84.80 \pm 2.51$ |
| ACGAN [20] | $87.45 \pm 1.33$ | $90.45 \pm 2.76$ | $84.57 \pm 1.39$ | $90.39 \pm 2.09$ |
| DAGAN | $89.93 \pm 0.88$ | $91.34 \pm 1.93$ | $88.57 \pm 0.90$ | $92.98 \pm 1.78$ |
| SCGAN$^\star$ | $94.26 \pm 0.63$ | $94.37 \pm 0.20$ | $94.89 \pm 0.53$ | $96.79 \pm 0.53$ |
| SCGAN | $94.30 \pm 0.48$ | $93.43 \pm 0.35$ | $95.14 \pm 0.42$ | $97.02 \pm 0.57$ |

We use both direct concatenation (i.e., DAGAN) and CASAM alignment (i.e., variant SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$) for the cross-domain fusion of text representation and images in $G_{s\to t}$, respectively. Compared with DAGAN, SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$ further improves the quantization performance, as shown in Tables 3 and 5, indicating that achieving alignment between domains in a brute force manner does not resolve the strong heterogeneity that exists between domains. DAGAN is essentially a pixel-level superposition of data from two different modalities. The mixing of data from different imaging principles affects the feature extractor's judgment on target data's feature distribution. Conversely, CASAM does not affect the independence of the feature distribution for each domain data. In particular, the semantic alignment layer can calculate the similarity between the generated image and the textual description before generating new image features. It can discover the semantic relationship between each pixel and words, mapping the image features to the corresponding fine-grained text representation.

Also, we quantitatively and qualitatively investigate the effects of $\mathcal{L}_{vd}$ and $\mathcal{L}_{CD-GP}$. Compared with SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$, SCGAN$-\mathcal{L}_{vd}$ adds a gradient penalty $\mathcal{L}_{CD-GP}$ to the discriminator to ensure the quality of the generated images. That is because $\mathcal{L}_{CD-GP}$ reduces the gradient of $I_t^{match}$ to the lowest point of the loss function curve while ensuring the smoothness of its adjacent regions, while other input images, such as $I_{s\to t}$, are placed on the high point of the curve. As shown in Fig. 8, the IS is significantly improved, indicating that $\mathcal{L}_{CD-GP}$ gives the generator a more explicit convergence target, guiding the generator to generate more realistic images and semantically consistent with ultrasonography reports. Further, our proposed SCGAN adds $\mathcal{L}_{vd}$ to learn discrepancy features. In principle, in our cross-modal domain adaptation task, the data of these two modalities are different in the visual layer but converge in the semantic layer. If the generator only learns the low-level visual layer features in the source domain, the prediction results mapped in the target domain will deviate from our expectations and penalize by the adversarial loss. However, the results of our SCGAN converge significantly,

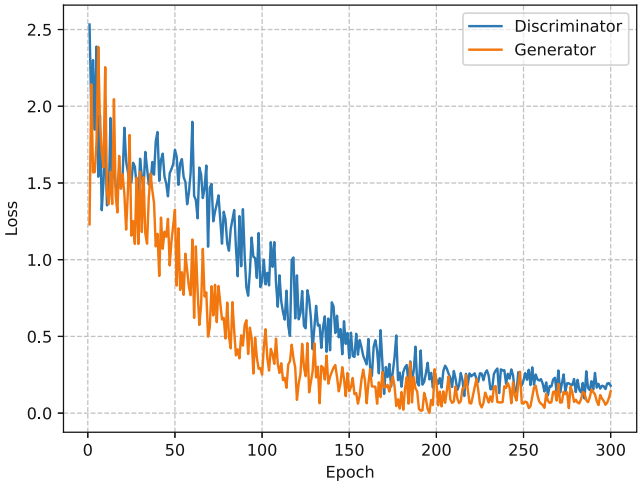


**Fig. 6** The loss curve of generator and discriminator in SCGAN

indicating that $G_{s\to t}$ learns high-level semantic layer features. Thus, $\mathcal{L}_{vd}$ can reverse encourage $G_{s\to t}$ to deceive $D_t$ in case of domain shifts, requiring $G_{s\to t}$ to capture high-level semantic domain invariant features across the source and target domains. As shown in Table 3, the accuracy and specificity are significantly improved, in Table 5, the IS is also boosted. The loss curves of the generator and discriminator in SCGAN are shown in Fig. 6. The experimental results prove the scientific validity of our techniques.

Specifically, we tune the parameters $\lambda_1$ and $\lambda_2$ in the loss function Equation (15), and the results are shown in Table 4 and Fig. 7. The IS significantly increases from 4.14 to 4.23 when $\lambda_2$ is changed from 0 to 2. Meanwhile, the IS increases to 4.26 when $\lambda_1$ is changed from 0 to 0.2, verifying the effectiveness of combining these two techniques. The IS score significantly decreased when $\lambda_2$ changed from 2 to 4 or $\lambda_1$ changed from 0.2 to 0.5. It may be that the penalty is too large, leading to the loss of some more important features. Therefore, in SCGAN, we set $\lambda_1$ and $\lambda_2$ to 0.2 and 2, respectively.

**Architecture Analysis** Table 5 reports the IS scores of SCGAN and other compared methods. We can observe that

**Table 3** Classification performance comparison of our SCGAN and its variants

| Methods | Results(%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| DAGAN | $89.93 \pm 0.88$ | $91.34 \pm 1.93$ | $88.57 \pm 0.90$ | $92.98 \pm 1.78$ |
| SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$ | $91.97 \pm 0.72$ | $92.23 \pm 0.83$ | $93.42 \pm 0.64$ | $95.35 \pm 0.74$ |
| SCGAN$-\mathcal{L}_{vd}$ | $93.72 \pm 0.43$ | $93.13 \pm 0.52$ | $94.57 \pm 0.57$ | $96.54 \pm 0.52$ |
| SCGAN | $94.30 \pm 0.48$ | $93.43 \pm 0.35$ | $95.14 \pm 0.42$ | $97.02 \pm 0.57$ |

**Table 4** Ablation study of loss function parameter adjustment results

| Results(%) | SCGAN($\lambda_1$, $\lambda_2$) | | | | |
|---|---|---|---|---|---|
| | 0, 0 | 0, 2 | 0, 4 | 0.2, 2 | 0.5, 2 |
| Inception Score | $4.14 \pm 0.23$ | $4.23 \pm 0.15$ | $3.97 \pm 0.30$ | $4.26 \pm 0.18$ | $4.21 \pm 0.37$ |

our model obtains the best score, significantly improving the IS from 2.58 to 4.26. Compared with DCGAN and ACGAN, we believe that the inclusion of domain information can guide the direction of the generator to generate images, which gives the generator has less freedom to generate images using random noise and reduces the uncertainty of the image generation process. In contrast, the multi-generator and multi-discriminator structures in StackGAN [48] and AttnGAN [42] make the quality of the generated images in the initial layer affect the final refinement, so the effect is poor. In conclusion, SCGAN can generate visually more realistic images with higher quality and better diversity than existing methods (Fig. 8).

In Table 6, we compare $\mathcal{L}_{vd}$ with the losses used in different methods. For example, SD-GAN [46] proposed a contrastive loss to improve the consistency between images generated by the same text description. Oord et al. [21] measured the dependence of two mutual information by learning the InfoNCE loss function and obtained a useful representation between the information. Wang et al. [39] used triplet loss to make video patches from the same trajectory closer in the embedding space than random patches. However, in contrastive loss and InfoNCE loss, all positive and negative matching pairs of each sample need to be sampled separately, and our $\mathcal{L}_{vd}$ does not need to dig the negative of information, which can reduce the complexity of training. Adding triplet loss to the baseline reduces the quality score of the generated image. This result shows that the better disentanglement of triplet loss may separate the connections between features too much and reduce the smoothness of interpolation.

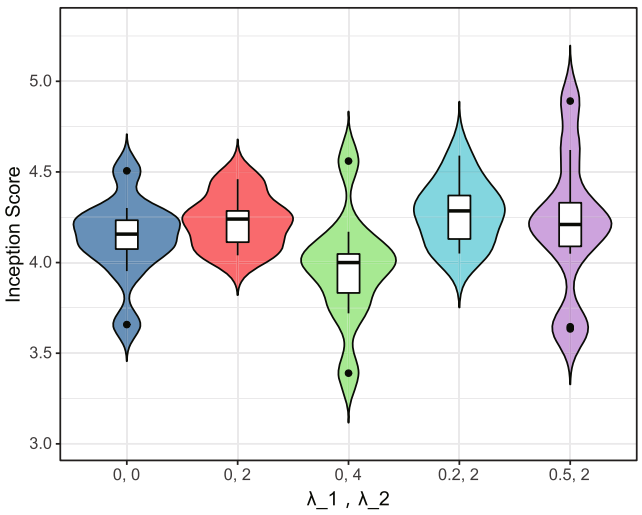


**Fig. 7** Inception score (IS) analysis for different parameters of the loss function

Table 7 gives the performance metrics of the classification models of SCGAN when pre-trained with VGGNet [29], GoogLeNet [30], ResNet-50, ResNet-101 and ResNet-152, respectively. The results show that the highest accuracy values are achieved using ResNet-50. Moreover, the trade-off between classification results and network optimization is crucial. Considering the dimensionality and parameter complexity of deep networks such as ResNet-101 and ResNet-152, and the relatively stable performance obtained with the ResNet series, we choose to use ResNet-50. Therefore, we use the modified ResNet-50 model to train our dataset and use it as a baseline classification method.

Figure 9 visualizes 24 images generated using DAGAN, SCGAN$-\mathcal{L}_{vd}$ $-$ $\mathcal{L}_{CD-GP}$ and SCGAN. Through human perception, we can find that compared with benign nodules, malignant nodules contain calcification (abnormal white spots) and irregular edge contours. From the perspective of the image quality generated, DAGAN without domain adaptation synthesizes nodules with irregular shapes, rough texture distribution, and lack of rich details. In contrast, the details of the nodules generated by CASAM gradually become clearer. However, the marginal area of some nodules changes greatly, which may be related to less marginal semantic information. The images generated by our SCGAN model are visually convincing. Among them, the internal grayscale difference of the nodules is obvious, and the tissue texture is clear. Benign nodules have smooth borders, and clustered calcifications accompany malignant nodules. It shows that the effective combination of CASAM and the two losses can potentially ensure the quality

**Table 5** The inception score (IS) of our proposed SCGAN and its variants compared with different image generation methods

| Results(%) | Inception Score | Results(%) | Inception Score |
|---|---|---|---|
| DCGAN [25] | $2.58 \pm 0.21$ | DAGAN | $3.48 \pm 0.37$ |
| ACGAN [20] | $3.01 \pm 0.22$ | SCGAN$-\mathcal{L}_{vd}-\mathcal{L}_{CD-GP}$ | $4.14 \pm 0.23$ |
| StackGAN [48] | $3.38 \pm 0.34$ | SCGAN$-\mathcal{L}_{vd}$ | $4.23 \pm 0.15$ |
| AttnGAN [42] | $3.47 \pm 0.24$ | SCGAN | $4.26 \pm 0.18$ |

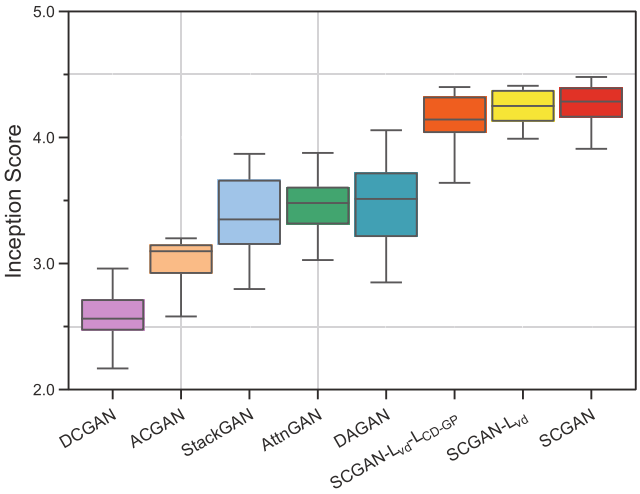


**Fig. 8** Box and whisker plot analysis of the inception score (IS) for different image generation methods

of the generated image, including the shape and texture distribution of the nodules.

**Physician Evaluation** We collaborate with three senior physicians who treat thyroid diseases. The whole process is divided into two parts. First, the three physicians independently determine the authenticity of the images, the semantic consistency, and the benignity or malignancy of the nodules and give their respective diagnoses. Then, in the second part, the three physicians could discuss and give the final results of the consultation. The accuracy of each physician and their mean values are shown in Table 8. Overall, our proposed model performs better than ultrasound physicians. The experiment results indicate that the highest individual score of the three physicians is 75.67%, and the consultation score is higher than the average value of the three physicians in determining the authenticity of the ultrasound images. For the diagnosis of benign and malignant nodules, the consultation score is higher than the three-person independent score, and its overall accuracy is higher than that of authenticity discrimination. We discuss further with the physicians and analyze the experimental results in detail. Physicians have a more accurate judgment of nodules with apparent

**Table 6** Compare the inception score (IS) of $\mathcal{L}_{vd}$ and other different losses

| Results(%) | Inception Score |
|---|---|
| Triplet Losses [39] | $3.92 \pm 0.18$ |
| Contrastive Loss [46] | $4.09 \pm 0.24$ |
| InfoNCE Loss [21] | $4.22 \pm 0.30$ |
| SCGAN$-\mathcal{L}_{vd}$ | $4.23 \pm 0.15$ |
| SCGAN | 4.26±0.18 |

**Table 7** Performance of our classification models of SCGAN when using pre-trained VGGNet, GoogLeNet, ResNet-50, ResNet-101, and ResNet-152, respectively

| CNNs | **Results(%)** | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| VGGNet [29] | 78.83 | 79.10 | 78.57 | 81.81 |
| GoogLeNet [30] | 81.75 | 83.58 | 80.00 | 84.95 |
| ResNet-50 | 84.67 | 88.06 | 81.43 | 88.28 |
| ResNet-101 | 83.21 | 86.57 | 80.00 | 86.93 |
| ResNet-152 | 82.48 | 85.70 | 80.00 | 86.14 |

benign or malignant features, such as thos with a regular shape, clear borders, or obvious calcification. In contrast, physicians need to observe cases over time in conjunction with review results, such as nodules with an irregular shape, blurred or uneven borders, and hypoechogenicity. The rate of misdiagnosis by physicians is higher when there are similar symptoms to thyroiditis or multiple endocrine adenomas. Therefore, physician consultation can provide more diagnostic experience for definitive classification results compared to individual judgment. The physicians also indicate that discriminating the authenticity of an image is more challenging than discriminating the benignity or malignancy of a nodule, which demonstrated the effectiveness of SCGAN's image generation capabilities. While discriminating the authenticity of the images, the physicians also evaluate the semantic consistency of the images. The results show that the image features match their associated knowledge descriptions, demonstrating the strength of our model in acquiring high-level semantic features.

**Comparison with State-of-the-Arts** Table 9 shows the performance comparison between the proposed SCGAN and nine state-of-the-art classification methods for thyroid nodules. The results show that the proposed model achieves a better classification performance. Since most of the datasets used for training models in the paper are derived from private datasets and the code is not open source, it is impossible to directly compare SCGAN with others' methods on the same datasets. Therefore, Table 9 lists the performance of these methods as recorded in the original published literature. Refs. [1, 23, 26] records the classification results of traditional methods. Refs. [36, 37, 50] records the classification results of deep learning methods under a single modality.The remaining three methods are similar to our proposed SCGAN in that they all choose to extract features from multimodal data. Among them, Yang et al. [43] and Qin et al. [24] both chose to extract features from images by fusing features from conventional ultrasound images with elasticity images.

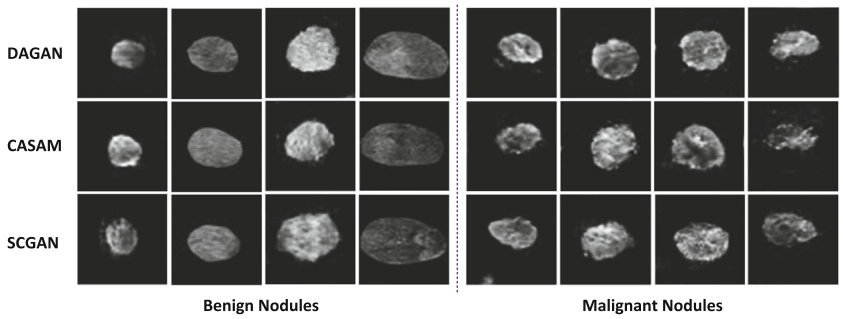


**Fig. 9** Visualization results of 24 images generated using DAGAN, CASAM (i.e., variant SCGAN$-\mathcal{L}_{vd} - \mathcal{L}_{CD-GP}$), and SCGAN

**Table 8** Evaluation results of three physicians on the authenticity of the images and the benignity and malignancy of the nodules

| Physician | Real/Generate Image | Benign/Malignant Nodule | | |
|---|---|---|---|---|
| | Accuracy | Accuracy | Sensitivity | Specificity |
| Physician 1 | 68.00 | 89.67 | 90.00 | 89.33 |
| Physician 2 | 75.67 | 91.33 | 90.67 | 92.00 |
| Physician 3 | 70.33 | 90.67 | 84.67 | 96.67 |
| Average of physicians | 71.33 | 90.56 | 88.45 | 92.67 |
| Consultation score | 73.67 | 93.67 | 93.33 | 94.00 |

**Table 9** Performance comparison of the SCGAN model with nine other existing methods for thyroid nodule classification

| Methods | Modality | Sample | Results(%) | | | |
|---|---|---|---|---|---|---|
| | | | Accuracy | Sensitivity | Specificity | AUC |
| Acharya et al. [1] | US(Texture features+C4.5) | 48 malignant, 223 benign | 94.30 | Not Given | Not Given | Not Given |
| Raghavendra et al. [26] | US(HOS + PSO +SVM) | 56 malignant, 288 benign | 97.71 | Not Given | Not Given | Not Given |
| Prochazka et al. [23] | US(histogram features +RF) | 20 malignant, 40 benign | 95.00 | 95.00 | 95.00 | 97.12 |
| Wang et al. [37] | US | 341 malignant, 705 benign | $87.32 \pm 0.0007$ | $84.22 \pm 0.0023$ | Not Given | $90.06 \pm 0.0007$ |
| Wang et al. [36]] | US | 524 malignant, 470 benign | 88.25 | 90.00 | 86.50 | 92.86 |
| Zhou et al. [50] | US | 1311 malignant, 4291 benign | Not Given | 98.70 | 98.80 | 98.00 |
| DScGANS [43] | US+USE | 1489 malignant, 1601 benign | $90.5 \pm 0.06$ | $88.1 \pm 0.08$ | $92.6 \pm 0.07$ | $91.4 \pm 0.04$ |
| Qin et al. [24] | US+USE | 617 malignant, 539 benign | $94.7 \pm 0.53$ | $92.77 \pm 1.04$ | $97.96 \pm 1.13$ | $98.77 \pm 1.05$ |
| KACGAN [28] | US+Text | 905 malignant, 1032 benign | $91.46 \pm 0.46$ | $90.63 \pm 0.38$ | $92.65 \pm 0.16$ | 95.32 |
| Proposed SCGAN | US+Text | 905 malignant, 1032 benign | $94.30 \pm 0.48$ | $93.43 \pm 0.35$ | $95.14 \pm 0.42$ | $97.02 \pm 0.57$ |

[a]Values are expressed as mean $\pm$ standard deviation

[b]Std is not provided in some sources

[c]US means ultrasound image, USE means ultrasound elasticity image

The former used information from different modalities to train DScGANS models to facilitate the diagnosis of benign and malignant thyroid nodules. The latter focused on comparing the effects of different fusion strategies and different classification network structures on classification performance. Compared to Qin, our method has higher sensitivity and similar accuracy, specificity, and AUC. However, all the above methods are constrained by the limited availability of annotated data. Differently, Shi et al. [28] instead used standardized terminology to assist in the extraction of ultrasound image features in KACGAN to facilitate thyroid nodule image enhancement. This method is similar to our idea, but our method does better in cross-modal alignment and obtains higher metric values. As mentioned above, cross-domain fusion using multimodal data to improve the classification performance of thyroid nodules has become a trend in thyroid nodule diagnosis.

## 5 Conclusion

In this paper, we propose a new deeply fused semantic consistency generative adversarial network (SCGAN) to diagnose benign and malignant nodules in thyroid ultrasound images. The method organically combines image features with textual information. The domain adaptation process of these two cross-modal data is accomplished jointly through the self-attention mechanism and metric learning, using their semantic consistency to reduce domain shifts in the training process. The addition of two new techniques to guide the hinge loss based on adversarial learning promotes the convergence of the network and improves the quality of image generation. The experimental results demonstrate the effectiveness of our SCGAN in improving the performance of target domain classification networks and have potential clinical applications.

We will work on a training model that can be applied to more types of ultrasound images and domain knowledge in future work. For example, the inclusion of richer knowledge information such as blood flow signals or ultrasound elasticity images improves diagnostic accuracy. Besides, the embedding process of our domain knowledge relies on pre-trained text encoders, which, unlike natural datasets, require parameter tuning for medical datasets. In the next step, we will add an attention mechanism to the text encoder to achieve the most advanced performance.

**Acknowledgements** This work is supported by the National Natural Science Foundation of China grant numbers 61972274, China Postdoctoral Science Foundation grant numbers 2018M631774 and Taiyuan 2019-nCoV prophylaxis and treatment research project grant numbers XG2020-5-04.

## Declarations

**Conflict of Interests** The authors declare that they have no conflict of interest.

## References

1. Acharya UR, Chowriappa P, Fujita H, Bhat S, Dua S, Koh JE, Eugene L, Kongmebhol P, Ng KH (2016) Thyroid lesion classification in 242 patient population using gabor transform features from high resolution ultrasound images. Knowl-Based Syst 107:235–245
2. Avola D, Cinque L, Fagioli A, Filetti S, Rodola E (2021) Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification. IEEE Trans Circ Syst Video Technol PP(99):1–1
3. Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis. In: International conference on learning representations
4. Bullinaria JA, Levy JP (2007) Extracting semantic representations from word co-occurrence statistics: a computational study. Behav Res Methods 39(3):510–526
5. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A (2017) Deep learning: a primer for radiologists. Radiographics 37(7):2113–2131
6. Chen C, Dou Q, Chen H, Heng PA (2018) Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In: International workshop on machine learning in medical imaging. Springer, pp 143–151
7. Chen C, Dou Q, Chen H, Qin J, Heng PA (2020) Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans Med Imaging 39(7):2494–2505
8. Chen J, You H, Li K (2020) A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Comput Methods Programs Biomed 185:105329
9. Dong H, Yu S, Wu C, Guo Y (2017) Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE international conference on computer vision, pp 5706–5714
10. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2016) Lstm: A search space odyssey. IEEE Trans Neural Netw Learn Syst 28(10):2222–2232
11. Gu Y, Ge Z, Bonnington CP, Zhou J (2019) Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. IEEE J Biomed Health Inform 24(5):1379–1393
12. 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
13. Kwon SW, Choi IJ, Kang JY, Jang WI, Lee GH, Lee MC (2020) Ultrasonographic thyroid nodule classification using a deep convolutional neural network with surgical pathology. J Digit Imaging 33(5):1202–1208
14. Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, Liu W, Zhang J (2018) An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Scientif Rep 8(1):1–12
15. Li Z, Yang K, Zhang L, Wei C, Yang P, Xu W (2020) Classification of thyroid nodules with stacked denoising sparse autoencoder. Int J Endocrinol 2020

16. Ma J, Wu F, Zhu J, Xu D, Kong D (2017) A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 73:221–230
17. Mescheder L, Geiger A, Nowozin S (2018) Which training methods for gans do actually converge? In: International conference on machine learning. PMLR, pp 3481–3490
18. Messina N, Falchi F, Esuli A, Amato G (2021) Transformer reasoning network for image-text matching and retrieval. In: 2020 25Th international conference on pattern recognition (ICPR). IEEE, pp 5222–5229
19. Moujahid D, Elharrouss O, Tairi H (2018) Visual object tracking via the local soft cosine similarity. Pattern Recogn Lett 110:79–85
20. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: ICML'17 Proceedings of the 34th International Conference on Machine Learning - vol 70, pp 2642–2651
21. Oord AVD, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv:1807.03748
22. Pan C, Huang J, Hao J, Gong J (2020) Towards zero-shot learning generalization via a cosine distance loss. Neurocomputing 381:167–176
23. Prochazka A, Gulati S, Holinka S, Smutek D (2019) Patch-based classification of thyroid nodules in ultrasound images using direction independent features extracted by two-threshold binary decomposition. Comput Med Imaging Graph 71:9–18
24. Qin P, Wu K, Hu Y, Zeng J, Chai X (2020) Diagnosis of benign and malignant thyroid nodules using combined conventional ultrasound and ultrasound elasticity imaging. IEEE J Biomed Health Inform 24(4):1028–1036
25. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR 2016 : International Conference on learning representations 2016
26. Raghavendra U, Gudigar A, Maithri M, Gertych A, Meiburger KM, Yeong CH, Madla C, Kongmebhol P, Molinari F, Ng KH et al (2018) Optimized multi-level elongated quinary patterns for the assessment of thyroid nodules in ultrasound images. Comput Biol Med 95:55–62
27. Ren J, Hacihaliloglu I, Singer EA, Foran DJ, Qi X (2018) Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 201–209
28. Shi G, Wang J, Qiang Y, Yang X, Zhao J, Hao R, Yang W, Du Q, Kazihise NGF (2020) Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. Comput Methods Prog Biomed 196:105611
29. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR 2015 : International Conference on learning representations 2015
30. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 1–9
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
32. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC et al (2017) Acr thyroid imaging, reporting and data system (ti-rads): white paper of the acr ti-rads committee. J Amer college Radiol 14(5):587–595
33. Ververas E, Zafeiriou S (2020) Slidergan: Synthesizing expressive face images by sliding 3d blendshape parameters. Int J Comput Vis 128(10):2629–2650
34. Wang H, Jia H, Lu L, Xia Y (2019) Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. IEEE J Biomed Health Inform 24(2):475–485
35. Wang H, Wang S, Qin Z, Zhang Y, Li R, Xia Y (2021) Triple attention learning for classification of 14 thoracic diseases using chest radiography. Med Image Anal 67:101846
36. Wang J, Li S, Song W, Qin H, Zhang B, Hao A (2018) Learning from weakly-labeled clinical data for automatic thyroid nodule classification in ultrasound images. In: 2018 25Th IEEE international conference on image processing (ICIP), pp 3114–3118
37. Wang L, Zhang L, Zhu M, Qi X, Yi Z (2020) Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Med Image Anal 61:101665
38. Wang M, Deng W (2018) Deep visual domain adaptation: a survey. Neurocomputing 312:135–153
39. Wang X, Gupta A (2015) Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802
40. Wang X, Li L, Ye W, Long M, Wang J (2019) Transferable attention for domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 5345–5352
41. Xie X, Niu J, Liu X, Chen Z, Tang S (2020) A survey on domain knowledge powered deep learning for medical image analysis. arXiv:2004.12150
42. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324
43. Yang W, Zhao J, Qiang Y, Yang X, Dong Y, Du Q, Shi G, Zia MB (2019) Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification. In: International conference on medical image computing and computer-assisted intervention, pp 558–566
44. Yao Y, Zhang Y, Li X, Ye Y (2019) Heterogeneous domain adaptation via soft transfer network. In: Proceedings of the 27th ACM international conference on multimedia, pp 1578–1586
45. Yi X, Walia E, Babyn P (2019) Generative adversarial network in medical imaging: a review. Medical image analysis 58:101552
46. Yin G, Liu B, Sheng L, Yu N, Wang X, Shao J (2019) Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2327–2336
47. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning. PMLR, pp 7354–7363
48. Zhang H, Xu T, Li H (2017) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: 2017 IEEE International conference on computer vision (ICCV), pp 5908–5916
49. Zhang Y, Wei Y, Wu Q, Zhao P, Niu S, Huang J, Tan M (2020) Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Trans Image Process 29:7834–7844
50. Zhou H, Wang K, Tian J (2020) Online transfer learning for differential diagnosis of benign and malignant thyroid nodules with ultrasound images. IEEE Trans Biomed Eng 67(10):2773–2780