UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

HT-RCM: Hashimoto's Thyroiditis Ultrasound Image Classification Model Based on Res-FCT and Res-CAM

Jiang, W.; Chen, K.; Liang, Z.; Luo, T.; Yue, G.; Zhao, Z.; Song, W.; Zhao, L.; Wen, J.

DOI
10.1109/JBHI.2023.3331944

Publication date
2024

Document Version
Final published version

Published in
IEEE Journal of Biomedical and Health Informatics

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Jiang, W., Chen, K., Liang, Z., Luo, T., Yue, G., Zhao, Z., Song, W., Zhao, L., & Wen, J. (2024). HT-RCM: Hashimoto's Thyroiditis Ultrasound Image Classification Model Based on Res-FCT and Res-CAM. *IEEE Journal of Biomedical and Health Informatics*, *28*(2), 941-951. https://doi.org/10.1109/JBHI.2023.3331944

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:24 Mar 2025

HT-RCM: Hashimoto's Thyroiditis Ultrasound Image Classification Model Based on Res-FCT and Res-CAM

Wenchao Jiang, Kang Chen, Zhipeng Liang, Tianchun Luo, Guanghui Yue, *Member, IEEE*, Zhiming Zhao, *Senior Member, IEEE*, Wei Song, Ling Zhao, and Jianxuan Wen

***Abstract*—The early lesions of Hashimoto's thyroiditis are inconspicuous, and the ultrasonic features of these early lesions are indistinguishable from other thyroid diseases. This paper proposes a Hashimoto Thyroiditis ultrasound image classification model HT-RCM which consists of a Residual Full Convolution Transformer (Res-FCT) model and a Residual Channel Attention Module (Res-CAM). To collect the low-order information caused by hypoechoic signals accurately, the residual connection is injected between FCTs to form Res-FCT which helps HT-RCM superimpose the low-order input information and high-order output information together. Res-FCT can make HT-RCM focus more on hypoechoic information while avoiding gradient dispersion. The initial feature map is inserted into Res-FCT again through a down-sampling component, which further helps HT-RCM exact multi-level original semantic information in the ultrasound image. Res-CAM is constructed by implementing a residual connection between a channel attention module and a convolution layer. Res-CAM can effectively increase the weights of the lesion channels while suppressing the weights of the noise channels, which makes HT-RCM focus more on the lesion regions. The experimental results on our collected dataset show that HT-RCM outperforms the mainstream models and obtains state-of-the-art performance in HT ultrasound image classification.**

***Index Terms*—Hashimoto's thyroiditis, ultrasound image, Res-FCT, Res-CAM, HT-RCM.**

Manuscript received 3 May 2023; revised 3 October 2023; accepted 2 November 2023. Date of publication 10 November 2023; date of current version 6 February 2024. This work was supported in part by the National Natural Scence Foundation of China under Grant 62072202, in part by Guangdong Natural Science Fund Project under Grant 2021A1515011243, in part by the EU H2020 CLARIFY (CLoud ARtificial Intelligence For pathologY) under Grant 860627. H2020-MSCA-ITN-2019 call, in part by National Natural Science Foundation of China under Grant 62371305 and 62001302, and in part by Guangdong Basic and Applied Basic Research Foundation under 2021A1515011348. *(Corresponding authors: Wenchao Jiang; Guanghui Yue; Jianxuan Wen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Ethics Committee of Guangdong Hospital of Traditional Chinese Medicine under Application No. ZE2023-214-01.

Wenchao Jiang, Kang Chen, Zhipeng Liang, and Tianchun Luo are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: jiangwenchao@gdut.edu.cn; chenkanghaoo@163.com; 320630010@qq.com; 544694396@qq.com).

Guanghui Yue is with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: yueguanghui@szu.edu.cn).

Zhiming Zhao is with the Multiscale Networked Systems research group, University of Amsterdam, 1098 XH Amsterdam, The Netherlands (e-mail: z.zhao@uva.nl).

Wei Song, Ling Zhao, and Jianxuan Wen are with the Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, Guangdong Province, Guangzhou, Guangdong Province 510120, China (e-mail: weiweiwindy@gzucm.edu.cn; lindazhao@gzucm.edu.cn; onesuanni@126.com).

Digital Object Identifier 10.1109/JBHI.2023.3331944

I. Introduction

HASHIMOTO'S Thyroiditis (HT) is an autoimmune thyroiditis characterized by infiltrating lymphocytes in the thyroid gland and existing thyroid autoantibodies in the serum [1]. The clinical manifestation of Hashimoto's thyroiditis is goiter, and the initial characteristics are not prominent which usually results in difficulty to be detected during the daily physical examination. And then, some symptoms such as hyperthyroidism, hypothyroidism, and even myxedema have occurred when visiting a doctor for the first time [2].

The diagnostic methods of HT include serological examination, pathological examination, and thyroid ultrasound examination and so on. Serum anti-thyroid peroxidase (TPOAb) is widely regarded as the best serological marker of HT. Investigations have shown that the specificity and sensitivity of elevated TPOAb in predicting HT are about 89.4% and 63.9% respectively [3]. Although nearly three of every four positive patients are conformed to have HT, the positive predictive accuracy of elevated TPOAb is still low. Fine needle puncture cytology is an important method for diagnosing of HT, but it has some limitations such as sampling error and unstable observation objects. In addition, fine needle puncture cytology cannot be routinely recommended in the clinical practice because it belongs to traumatic examination. Therefore, utilizing noninvasive imaging technology has become a significant way for clinical diagnosis of HT in recent years [4].

The AI technology shows a powerful ability of feature representation, and it has been proved that AI can achieve good recognition performance for medical image classification [5]. So, AI-assisted ultrasound image processing is a feasible diagnosis method for intelligent medical treatment [6]. Nevertheless, the ultrasonic features of HT are variable and indistinguishable from

2168-2194 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

other thyroid diseases [7], [8], [9]. Furthermore, the symptoms response of HT in different stages also changes, and objective differences exist in different patients [10], [11], [12]. The traditional CNN model with a small convolution kernel is difficult to capture both the long-distance semantic information and the global semantic information due to its limited receptive field, and so can not effectively learn simultaneously both the detailed features and the structural information of the HT ultrasound image [13], [14]. Therefore, considering the characteristics of HT ultrasound image such as unclear lesion features, a large area of the hypoechoic region, unclear boundary, and complex texture features, it is essential to study a neural network model which can makes full use of both the long-distance feature information and the global semantic information of the HT ultrasound image.

In this paper, a HT Residual Classification Model (HT-RCM) is proposed based on the Residual Full Convolution Transformer (Res-FCT) model and the Residual Channel Attention Module (Res-CAM). The Res-FCT model mainly helps HT-RCM to extract the hypoechoic information, and the Res-CAM mainly helps HT-RCM to focus more on the lesion regions. HT-RCM can effectively recognize the lesion characteristics existing in the HT ultrasound image and improve the detection accuracy of early lesions of HT. The main contributions of this paper are as follows:

1) The Res-FCT model is designed by injecting the residual connection between Fully Convolutional Transformer (FCT) to extract the low-order information produced by hypoechoic signals. Res-FCT can superimpose low-order input information and high-order output information, making HT-RCM always focus on low-order information while avoiding diffusion of gradients and information loss. In addition, to make HT-RCM learn multi-level original semantic information, a down-sampling component is continuously conducted on the initial feature map and is then injected into the corresponding Res-FCT to further improve the learning ability of HT-RCM.
2) The Res-CAM is designed by implementing a residual connection between the channel attention module (CAM) and the convolution layer to make HT-RCM to focus more on the lesion regions where the visual features are usually weaker than those in other noise regions. Res-CAM can increase the weights of the lesion channels while suppressing the weights of the noise channels, helping to extract the characteristics of the lesion region more accurately where there is noticeable noise around it, and improving classification accuracy and generalization ability of HT-RCM.
3) An HT ultrasound image dataset is constructed and used to verify the effectiveness of the HT-RCM model. This ultrasound image dataset is composed of 8000 images collected from the Endocrinology Department of four branch hospitals in Guangdong Provincial Hospital of Chinese Medicine in the last 2 years. Extensive experiments show that the HT-RCM model can achieve state-of-the-art performance in HT ultrasound image classification.

II. Related Work

A. AI-Assisted Ultrasound Diagnosis

Ultrasound is cheap, high resolution and easy to operate [15], but ultrasound image has the characteristics of high noise and low resolution. There are also differences between ultrasound images generated by different imaging devices which poses great challenges for doctors to make accurate diagnosis. AI technology can effectively learn and recognize the feature information of ultrasound images, and it has been widely used to diagnose many diseases in the past decade. Hsieh et al. created a CKD (Chronic Kidney Disease) analysis system based on a support vector machine to classify and predict chronic kidney disease through ultrasound image, and the system's specificity reached 83.34% [16]. To assist doctors in diagnosing benign or malignant thyroid nodules, Shi et al. used ResAt-Faster R-CNN model to detect and classify thyroid nodule lesions in ultrasound image [17]. In order to reduce the burden of data annotation, Song et al. proposed an automatic unsupervised learning method based on double reconstruction, which can accurately recognize and classify breast tumor through ultrasound image [18]. In order to identify benign and malignant breast cancer, Zhao et al. developed a CAD (Computer-Aided Diagnosis) network model based on MAFS (Multi-Angle Fusion Strategy) and achieved good results in ultrasound image recognition for breast lesions [19]. Li et al. used a deep convolutional neural network model for feature learning of thyroid ultrasound image, and performed better than skilled radiologists in thyroid cancer recognition [20]. Gao et al. developed a two-branch combined network (DCN) for COVID-19 diagnosis which can effectively improve the accuracy of COVID-19 diagnosis [21], and the maximum classification accuracy of DCN on the validated dataset reached 96.74%. In order to make full use of the global features of ultrasound images, Zhang et al. proposed a multi-stage hybrid Transformer (MSHT) to combine the advantages of CNN and Transformer. The classification accuracy of MSHT in 4240 ROSE image datasets reached 95.68% [22]. Sheng et al. proposed a dual-attention deep manifold harmonic discrimination method for early diagnosis of neurodegenerative disease, and achieved excellent classification accuracy and robustness on two independent datasets [23].

B. Ultrasound Diagnosis of HT

Due to the unobvious lesion features and large area of noise region existing in the HT ultrasound image, it is difficult for AI model to fully extract the texture information of the ultrasound image. To solve this problem, Acharya et al. designed a computer-aided diagnosis model using gray features to classify HT. The model extracted gray features from ultrasound image of HT and constructed a support classifier based on the obtained feature vectors to detect HT, and obtained an accuracy up to 80% [24]. In addition, Acharya et al. proposed a new thyroid scanning model which can capture the texture features of thyroid ultrasound image to improve the accuracy of the diagnosis of HT [25]. Zhang et al. proposed an HTNet classification model

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

based on the residual network to predict HT and achieve good results [26]. However, the scope of learning features of the convolutional layer in the residual network was limited, resulting in a failure to capture and learn the long-distance semantic information and global semantic information in the ultrasound image. In order to improve the diagnostic accuracy of HT, Kang et al. integrated nine CNNs into a CAD-HT model which can improve the diagnostic accuracy effectively [27]. However, the specific size of the convolution kernel in the CAD-HT model limited the receptive field of the model, which made it also lacking the ability of learning both long-distance semantic information and global semantic information. In addition Liang et al. proposed an HT ultrasound image classification model HTC-Net based on residual network reinforced by channel attention mechanism. HTC-Net obtained high performance in early lesion recognition in HT ultrasound image [28]. However, HTC-Net was developed based on ResNet, and its receptive field was limited and cannot capture the long-distance semantic information of the image well. Due to the fact that traditional CNN models cannot fully extract the texture information of HT ultrasound images, and the generalization ability is also poor [29]. So, we need to research a more effective neural network model for HT ultrasound image recognition and classification.

C. Attention Mechanism in AI

When people pay attention to a thing, they usually first observe it roughly and then focus on the key feature points of the thing, so as to quickly and accurately understand its basic features. To mimic the way that human learn and recognize things, some works have been conducted and make the AI model focusing more on the target region by using an attention module. In order to overcome the inefficiency of existing transform based models to understand the fine-grained nature of segmentation tasks, Tragakis et al. proposed a Fully Convolutional Transformer with an attention module which can effectively extract the medium or long-distance semantic information and meet the needs of the refined processing of medical image analysis [30]. Woo et al. proposed a Convolutional Block Attention Module (CBAM) which can effectively improve the performance of different models in classification and detection tasks [31]. In order to consolidate multi-scale aggregation while learning channel attention more efficiently, Bark et al. proposed an efficient multi-scale channel attention module which improved the accuracy of ResNet 18, 34 and 50 ImageNet benchmarks by 0.8%, 0.6% and 1% respectively compared to SENet [32]. Considering that CNNs do not explicitly model the associations or relative importance of features in the spectral/temporal/channel-wise axes, Guo et al. proposed an spectral-temporal-channel (STC) attention module [33].

III. Methods

A. Overall Framework

HT-RCM model is composed of a Res-FCT model and a Res-CAM module. The Res-FCT model can effectively learn both the global semantic information and the long-distance semantic information in ultrasound image. The Res-CAM module makes the HT-RCM model to focus more on the lesion regions and reduces the interference of the noise regions. Under the cooperations of these two components, HT-RCM can fully extract global texture features, global semantic information and long-distance semantic information of the lesion regions existing in the ultrasound image, while obtaining stronger generalization ability at the same time. The overall framework of HT-RCM is shown in Fig. 1.

B. Res-FCT

Traditional CNN uses convolution kernels with a specific size to extract image features, but the convolution kernels with fixed size make the model to focus more only on the local features of the image [34]. So, a large amount of global semantic information and long-distance feature information in the ultrasound image is missing which actually is important to recognize the HT lesions [35]. In order to solve the above problems, the Res-FCT model is constructed through connecting multiple FCTs with residual connections [36], and then a down-sampling component is continuously conducted on the initial feature map and injected into the Res-FCT model to further improve the learning ability of HT-RCM. Different from the original residual connection in the FCT which can be understood as intra-link inside the same FCT, the residual connections of Res-FCT can be understood as inter-links added between different FCTs. The residual connection in Res-FCT superimposes the low-order input information and the high-order output information to help the model pay more attention to the low-order information. In addition, the Res-FCT model can avoid performance degradation through the effective information superimposing [37]. The structure of the Res-FCT model is shown in Fig. 1(a).

Generally, the texture information of HT ultrasound image is complex, and directly using the raw ultrasound image for model training may produce large amount of training parameters and make the model converge slowly. To solve this problem, a convolution layer composed of 7×7 convolution kernels is inserted before the FCT to reduce the image size and increase the number of the channels. In addition, the BatchNormal layer is inserted after the convolution layer to normalize the feature map and prevent the over-fitting of the model. Then the ReLU activation function is applied to the output of the BatchNormal layer, and the ReLU can enable the model to learn the target features better and fit the training data by activating most neurons simultaneously. In order to further reduce the number of the training parameters, the Max pooling operation is used on the feature map following the ReLU activation function to compress the feature map size. The specific operation implemented on FCT is shown in (1):

$$X_1 = \mathrm{MaxPool}(BN(\mathrm{Conv}(X))) \tag{1}$$

where X represents the initial input feature map of the Res-FCT model and X_1 represents the input feature map of the MaxPool preprocessing of X. In the FCT, X_1 undergoes LN layer normalization, i.e. LayerNorm, operation, convolution operation and Max pooling operation in turn to obtain the feature map

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

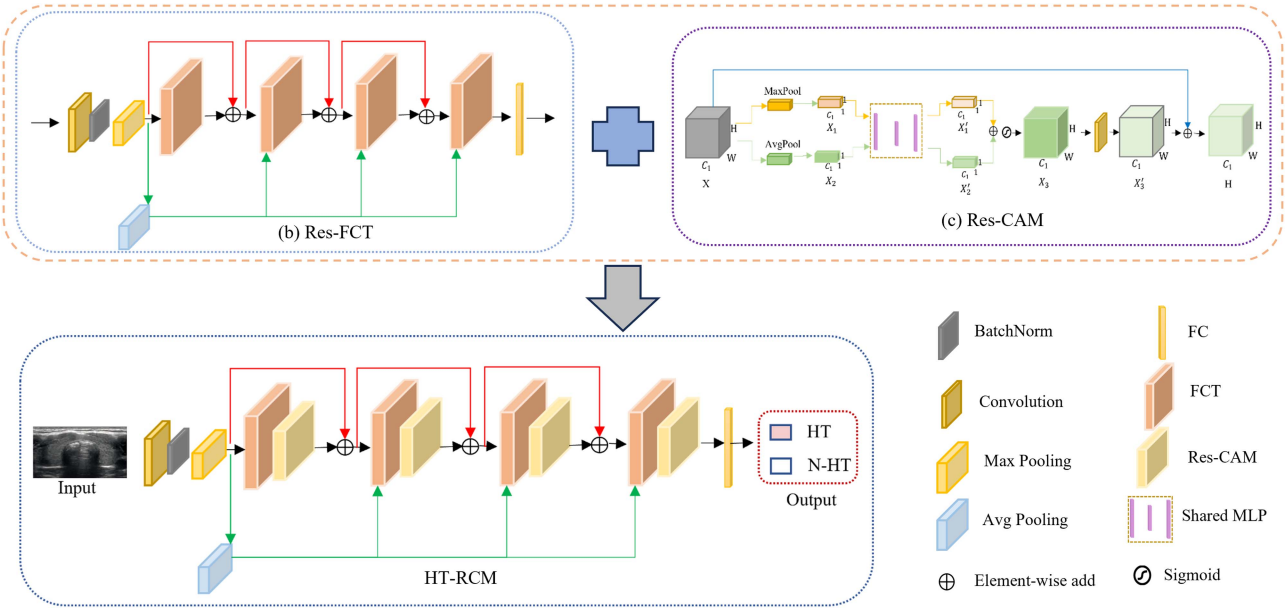

Fig. 1. Framework of HT-RCM. HT-RCM consists of the Res-FCT model and Res-CAM, and the structure of Res-FCT and Res-CAM are shown in (a) and (b) respectively. In (a), we first concatenate multiple FCT modules, and then overlay the inputs and outputs of the different FCT modules through residual connections, shown as red solid lines in (a), to help the model to focus more on low-order features. Furthermore, to further enhance the learning ability to the multi-scale information, we continuously conduct down-sampling, shown as green solid lines in (a), to obtain the feature map through the average pooling layer and inject it into the FCT module again. In (b), we add a convolution layer after the Sigmoid operation to further extract the lesion features, and use the residual connection to superimpose input and output to enhance the focusing ability to extract lower-order features of the ultrasound image.

X_1'. In order to enhance the model's learning ability to the long-distance semantic information, the convolution attention module is used to process the feature image X_1'. The convolution attention module can effectively capture and learn the long-distance semantic information of the image through depthwise convolution operation. The output y_0 of convolution attention block is expressed as (2) and (3):

$$y_0 = \text{MHSA}\left(X_1'\right) + X_1^{q/k/v} \tag{2}$$

$$X_1^{q/k/v} = Flatten(DepthConv(Reshape(X_1))) \tag{3}$$

where $X_1^{q/k/v}$ represents the operation vector of q, k, and v generated by the depthwise convolution operation on the feature map, and MHSA represents the Multi-Head Attention mechanism. In the wide-focus module of FCT, y_0 firstly goes through the LN layer for normalization operation to obtain the feature map. Then it enters the dilated convolution layers with convolution expansion rates of 1, 2, and 3 respectively to capture the semantic information with different sizes of receptive fields under different convolution expansion rates. Furthermore, the semantic information captured by the three convolutional layers is superimposed to generate the global semantic information y_1 so that the Res-FCT model can learn multi-level semantic information. Finally, the convolved y_1 and y_0 are summed as the output of the Wide-Focus module. The output expression of the first FCT module is shown as (4):

$$FCT(X_1) = MaxPool(Conv(y_1) + y_0)) \tag{4}$$

In order to further enhance the learning ability of the model for the low-order semantic information, avoid degradation caused by too deep network, reduce complexity, and speed up training speed, the residual connection is inserted between multiple FCT modules. The residual connections can help the network superimpose low-order input information and high-order output information. Thus, the problem of missing original feature caused by too deep network structure is solved. In addition, to further enhance the model's ability to learn global semantic information, the Res-FCT model uses the pyramid form to down-sampling the original feature map and injects it into the Res-FCT module again, and then the concat operation is used to fuse the down-sampled original feature map and the input feature map. In this way, the channel of the Res-FCT model can be widen and learn richer feature information of the lesions. The expressions of the output of the second, third, and fourth FCT modules in Res-FCT are shown as (5)–(7) respectively:

$$y_2 = FCT(Concat((FCT(X_1) + X_1), downsampling(X_1))) \tag{5}$$

$$y_3 = FCT(Concat((y_2 + FCT(X_1) + X_1), downsampling(X_1))) \tag{6}$$

$$y_4 = FCT(Concat((y_3 + y_2 + FCT(X_1) + X_1), downsampling(X_1))) \tag{7}$$

where Concat represents vector concatenation.

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

C. Res-CAM

The Channel Attention Module (CAM) can make the AI model to focus on the meaningful information by compressing the spatial dimension and preserving the channel dimension of the feature map. However, the texture features of the lesion region in HT ultrasound image are not evident. The noise region, sometimes existing as thyroid nodules, is larger than the lesion region. Therefore, the differences of the weights between the noise feature channels and the lesion feature channels are still not evident enough after the feature map is processed only by the CAM. The testing shows that using only CAM still pays too much attention to the noise region. To solve this problem, a Residual Channel Attention Module (Res-CAM) is constructed as shown in Fig. 1(b). The Res-CAM inserts a convolution layer to extract the lesion features of the output map to enhance the lesion features and weaken the noise features. Meanwhile, both the input and the output of Res-CAM are superimposed using a residual connection to enhance the weights of lesion feature channels and weaken the weights of the noise feature channels. The Res-CAM further improves the lesion recognition ability of HT-RCM for HT ultrasound image classification.

In Res-CAM, the input feature map X passes through the Max pooling layer and the Average pooling layer simultaneously, and the dimension of the feature map $H \times W \times C_1$ becomes $1 \times 1 \times C_1$. The detailed expressions are shown in (8) and (9):

$$X \in \mathbb{R}^{H\times W\times C_1} \xrightarrow{\text{MaxPool}} X_1 \in \mathbb{R}^{1\times 1\times C_1} \tag{8}$$

$$X \in \mathbb{R}^{H\times W\times C_1} \xrightarrow{\text{AvgPool}} X_2 \in \mathbb{R}^{1\times 1\times C_1} \tag{9}$$

where H and W represent the height and width of the input feature map respectively, and C_1 represents the number of channels. MaxPool and AvgPool can reduce the feature map size, X_1 and X_2 are processed by the shared MLP module, resulting X_1' and X_2' respectively. The purpose of the shared MLP module is compressing the channel number of the feature map into C_1/r and then extending the channel number into C_1, thereby the model can effectively filter noise information and extract meaningful lesion features. The detailed expression of the Shared MLP module is shown as (10) and (11):

$$X \in \mathbb{R}^{1\times 1\times C_1} \xrightarrow{\frac{1}{r}} X \in \mathbb{R}^{1\times 1\times \frac{C_1}{r}} \tag{10}$$

$$X \in \mathbb{R}^{1\times 1\times \frac{C_1}{r}} \xrightarrow{r} X \in \mathbb{R}^{1\times 1\times C_1} \tag{11}$$

where r represents the channel compression ratio. In order to improve the model sparsity, the ReLU activation function is used to process X_1' and X_2', and then the processed X_1' and X_2' are summed to obtain X_3. In this way, the weights of both the strong feature channel and the background information channel are summed together to generate a new feature map. The Sigmoid activation function is used to normalize X_3 and multiply it with the feature map X to enhance the strong feature information retaining the background information of the initial image. Finally, the convolution layer is used to further extract important features.

Besides, the weights of the lesion feature channels from the input feature map are larger than that from the noise feature channels in Res-CAM. Therefore, superimposing the output and input of the Res-CAM using residual connection can enhance the weights of lesion feature channels and weaken the weights of noise feature channels. The difference of the weights between the lesion feature channels and the noise feature channels can be enlarged effectively. Meanwhile, the residual connection can enable the model to capture low-order information and avoid performance degradation caused by information loss. The above process of Res-CAM is shown as (12):

$$H(X) = Conv(\sigma(MLP(AvgPool(X)) + MLP(MaxPool(X)))) + X \tag{12}$$

where σ represents sigmoid function, and MLP represents multi-layer perceptron.

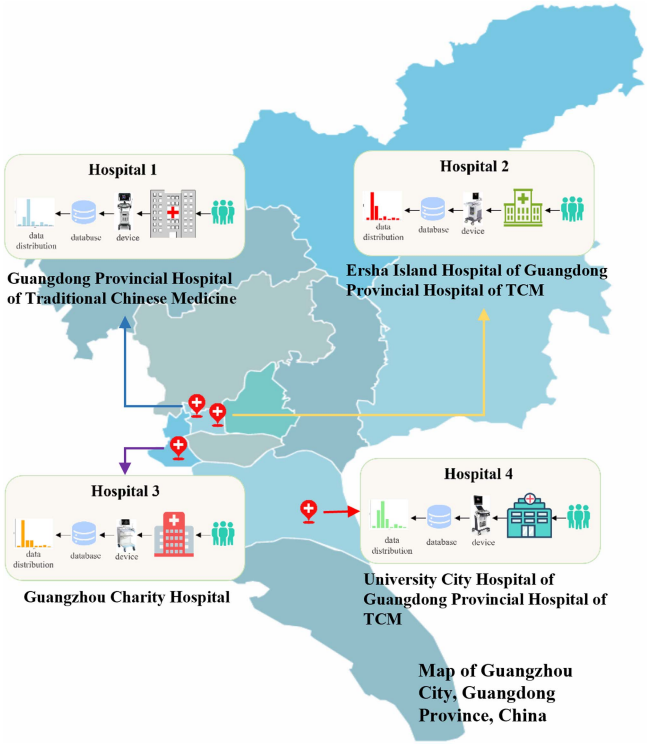

Fig. 2. Data source diagram. Hospital 1 is the Guangdong provincial hospital of traditional Chinese medicine, hospital 2 is Guangzhou charity hospital, hospital 3 is the university city hospital of Guangdong provincial hospital of TCM, and hospital 4 is the Ersha Island hospital of Guangdong provincial hospital of TCM. The ultrasound devices used in hospital 1 and 3 are GE LOGIQ E9, and the ultrasound devices used in hospital 2 and 4 are Canon i900.

IV. Experimental Results and Analysis

A. Dataset

The experimental dataset is constructed based on 8,000 HT ultrasound images collected by the Endocrinology departments of four branch hospitals of Guangdong Hospital of Traditional Chinese Medicine between January 2020 and December 2022. The data source diagram and the detailed statistical information of the dataset are shown in Fig. 2 and Table I. The dataset

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DETAILED STATISTICS OF THE DATASET

Hospital	Nums	Age(Range)	Sex(Male:Female)	Device
Hospital 1	2036	16-85	0.61:4.51	GE LOGIQ E9
Hospital 2	2021	18-80	0.62:5.23	Canon i900
Hospital 3	1953	17-82	0.53:5.16	GE LOGIQ E9
Hospital 4	1990	20-80	0.46:5.12	Canon i900

is randomly divided into two groups by 9:1 as training set and validation set respectively. The use of the dataset for related research has been approved by the Ethics Committee of Guangdong Hospital of Traditional Chinese Medicine (Number: ZE2023-214-01, Date: 21 June 2023).

B. Experimental Environments

The experimental environments in this paper are GPU3090, Pytorch1.12.1 and Numpy1.21.5. In the experiments, the model is trained for 250 epochs each time, and the average of 10 training is taken as the final statistical results. During the training phase, the initial image is set to $224 \times 224 \times 3$, Batch Size is set to 64, the learning rate is $2 \times e^{-3}$, and the optimizer is set to Adam. In addition, to increase the diversity of the dataset, the images in the training set are translated, randomly rotated or changed color to improve the model's generalization ability.

C. Evaluation Metrics

In order to evaluate the classification performance of different models for the HT ultrasound image dataset, the following four evaluation metrics which are Sensitivity, Specificity, Precision and F1-score are used:

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - score = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision} \tag{16}$$

TP (True Positive) means that a sample is classified as a positive sample and it is actually a positive sample, FP (False Positive) means that a sample is classified as a positive sample but it is actually a negative sample, TN (True Negative) means that a sample is classified as a negative sample and it is actually a negative sample, and FN (False Negative) means that a sample is classified as negative but it is actually a positive sample. Sensitivity represents the proportion of the correctly predicted positive samples in the total positive samples, specificity represents the proportion of correctly predicted negative samples in the total negative samples, and F1-score is used to measure the accuracy of the model.

D. Compared Method

We compared HT-RCM with commonly used CNNs and HT-related models as shown in Table IV. Among them, EfficientNetV2 is an improved CNN based on the EfficientNet series, achieving efficient feature learning through uniform scaling in terms of network depth, width, and resolution [38]. ResNet50 is composed of multiple residual blocks, effectively addressing gradient vanishing and exploding issues during training [36]. DenseNet is a deep CNN that provides dense connections and feature reuse, increasing network depth and enhancing feature extraction capabilities [39]. AlexNet is a classic deep CNN known for its strong performance on large-scale datasets [40]. These four CNNs have made significant contributions in medical image classification. For example, Jalehi et al. proposed a deep CNN model based on EfficientNetV2, achieving a sensitivity of 98.66%, specificity of 99.51%, and accuracy of 99.4% on CXR dataset tests [41]. Huang et al. introduced the ECA-EfficientNetV2 model based on EfficientNetV2, achieving an accuracy of 99.81% on a chest CT dataset, demonstrating excellent performance [42]. Panthakkan et al. developed the X-R50 model based on Xception and ResNet50, with a predictive accuracy of 97.8% on 10,500 skin images, showcasing high accuracy and effectiveness [43]. To overcome the limitations of existing models, Prakash et al. proposed a deep liver abnormality detection method based on DenseNet, achieving an accuracy of 98.34%, outperforming other pioneer methodologies [44]. Eldem et al. experimented with six different variants of the AlexNet architecture, where 6Conv_SVM was tested on diabetic and pressure wound images in the public medetec dataset, achieving an accuracy of 95.33% and demonstrating strong performance [45]. Therefore, considering the contributions of ResNet, EfficientNetV2, DenseNet, and AlexNet in medical image classification, as well as the consideration of model size, we have chosen ResNet50, EfficientNetV2_S, DenseNet, and AlexNet as the comparative CNN models. In addition, due to lacking directly relevant works and the open-source implementations on HT ultrasound image classification, we have compare ResNet152 and HTC-Net in our comparative experiments. However, as HTNet [26] combines serological markers and image data for HT classification, it cannot be directly used in our experiments. Therefore, we have only included the base model of HTNet, ResNet152, in our comparative experiments.

E. Impact of the Attention Module

There are different kinds of attention modules such as Channel Attention Module (CAM), Spatial Attention Module (SAM) and Convolutional Block Attention Module (CBAM) consisting of both CAM and SAM [31]. Among them, CAM makes the model focus on the target information by keeping the channel dimension unchanged and compressing the spatial dimension. SAM makes the model focus on the target location information by keeping the spatial dimension unchanged and compressing the channel dimension. Any kind of attention modules mentioned above can help the AI model focus on strong feature regions. However, we actually select CAM in HT-RCM because

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
OPTIMAL SOLUTION OF ATTENTION MECHANISM

Model	Accuracy(%)
Res-FCT-SAM	78.98
Res-FCT-CBAM	81.53
Res-FCT	82.73
Res-FCT-CAM	**83.33**

The bold sections indicate that the individual models that performed best in the experiment are highlighted in bold.

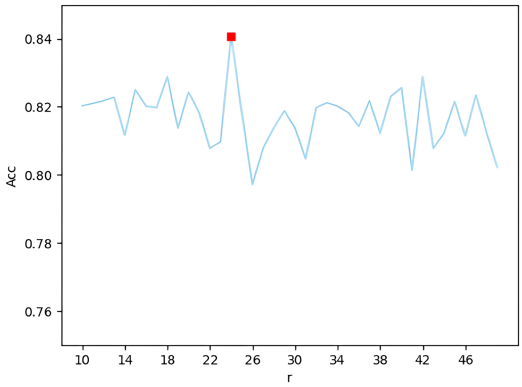

Fig. 3. Optimal solution for *r* parameter. We can find that the channel compression ratio *r* is in the range of 10 to 50, and Res-CAM has the best effect when $r = 24$.

the real experimental results shows that CAM is more effective than others. To analyze the focusing ability of CBAM, CAM and SAM to the lesion regions in HT ultrasound image, some parameter tuning testing and comparison are implemented based on the backbone model Res-FCT, and the experimental results are shown in Table II.

Table II shows that the classification accuracy of Res-FCT-SAM (Res-FCT + SAM), Res-FCT-CBAM (Res-FCT + CBAM), Res-FCT, and Res-FCT-CAM (Res-FCT + CAM) are 78.98%, 81.53%, 82.73%, and 83.33% respectively. The Res-FCT-SAM obtains the lowest classification accuracy which indicates that the SAM inhibits the model's ability to extract lesion features. So, AI models with SAM are not suitable for the lesion recognition of HT ultrasound images. The classification accuracy of Res-FCT-CAM is the highest among all attention modules which indicates that the CAM can effectively improve the model's attention to the lesion region, and so can enhance the model's recognition and classification ability for HT ultrasound image.

In addition, the channel compression ratio *r*, as mentioned in (10) and (11), of Res-CAM is an adjustable and key parameter which can affect the performance of the HT-RCM model. In order to select the optimal value of *r*, experiments are carried out and the experimental results are shown in Fig. 3. As can be seen from Fig. 3, when *r* is located in the range of 10 to 23, the classification accuracy of HT-RCM fluctuates up and down with a small amplitude. When $r = 24$, the classification accuracy of HT-RCM is increased significantly and reaches the peak value. However, when *r* locates in the range of 25 to 50, the classification accuracy of HT-RCM decreases significantly and then float up or down with a small amplitude. Therefore, when $r = 24$, the HT-RCM achieves the optimal lesion recognition and classification efficiency for HT ultrasound image.

F. Grad-CAM Experiments

In order to analyze and compare the focusing performance of different models for HT ultrasound images during the training stage, the experiments adopt Grad-CAM for visualization. The Grad-CAM is calculated according to the feature map *A* of the last convolutional layer after training [46]. The final partial derivative feature map *G* is obtained by calculating the partial derivatives of all pixels in each channel of the input feature map *A*. Then, the global average pooling (GAP) is implemented on the partial derivative feature map *G* to obtain the weight vector *r* composed of *M* elements. The expression of the weight vector *r* is shown in (17):

$$r = [r_1, \ldots, r_k, \ldots, r_M] \tag{17}$$

where *M* represents the number of the channels of *G*, r_k represents the average sensitivity of the target category to the *k* channel from the last layer of feature map. The expression of r_k is given in (18):

$$r_k = \frac{1}{HW} \sum_{i}^{H} \sum_{j}^{W} G_{i,j}^{k} \tag{18}$$

where *H* and *W* represent the height and width of feature map A respectively, $G_{i,j}^{k} = \frac{\partial y^b}{\partial A_{i,j}^{k}}$, y^b represents the predicted score of the model for category b, and $A_{i,j}^{k}$ represents the point with coordinate *(i, j)* in the *k* channel of the feature map *A*. Finally, the weight vector *r* and the corresponding channel of the feature figure *A* are linearly weighted to obtain a visualization map. The specific expression of linear weighting is shown in (19):

$$GradCAM = ReLU \sum_{k} r_k A^k \tag{19}$$

Specially, to evaluate the efficiency of the Res-CAM, we compare the visualization results of Res-CAM and other channel attention methods using Grad-CAM, and the results are shown in Fig. 4. From Fig. 4, we can see that SENet pays more attention to the noise region, CAM and ECA pay attention to the lesion region while still paying attention to large-scale noise region. Res-CAM can focus on the lesion region in the ultrasound image more accurately and help the model obtain most accurate feature representing ability. So we select Res-CAM as the attention module in the experiments of HT-RCM model. Overall, Fig. 5 shows the Grad-CAM visualization results our HT-RCM using Res-CAM and other classification models. The last layer feature map is selected for display in the experiments. In the thermal diagram, brighter color (the redder the color is, the greater the brightness is) indicates that the model pays more attention to this area. The area included in the red box represents the actual lesion region in the HT ultrasound image.

In Fig. 5(a)–(c) represent the raw HT ultrasound images respectively, and *(a(x))*, *b(x)*, and *c(x) (x = 1, 2,..., 5)* represent

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

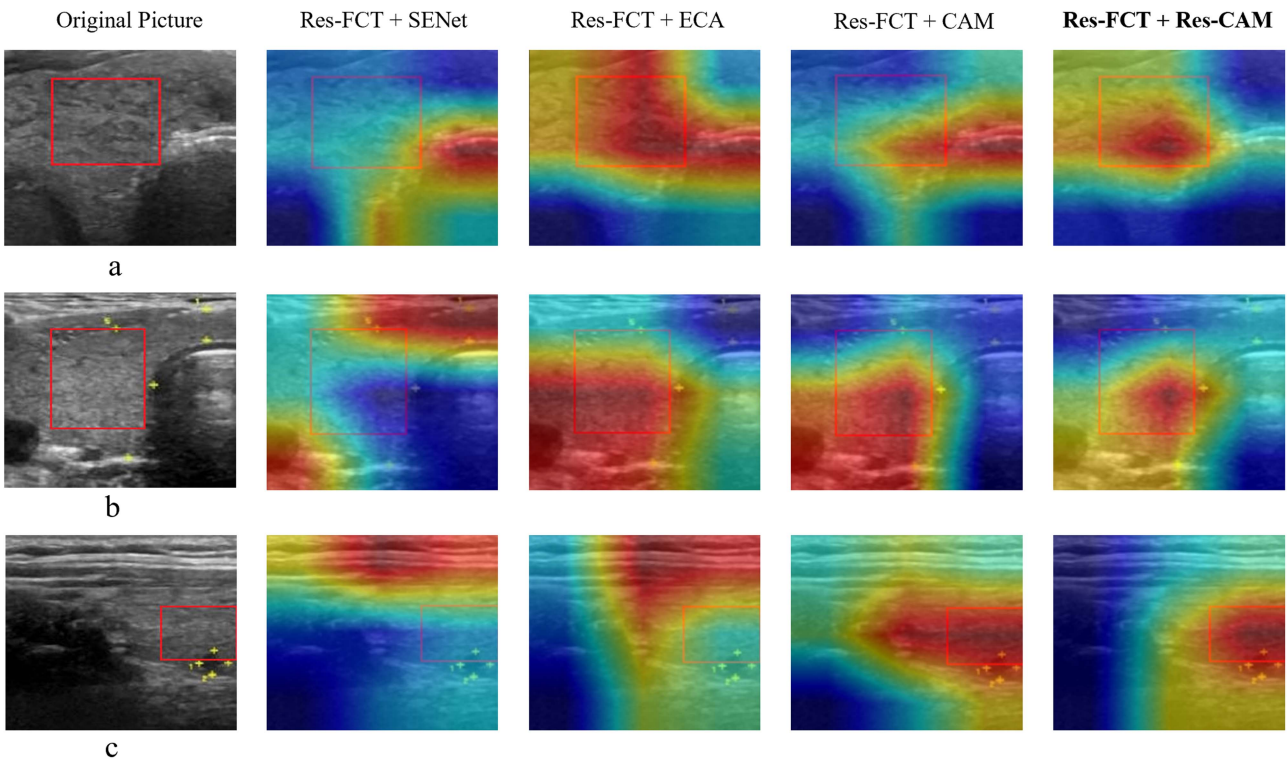

Fig. 4. Visualization results of Grad-CAM experiments. We compare the visualization results of different attention modules, i.e. SENet, ECA, CAM, and Res-CAM. For convenience, we add the name of backbone network, Res-FCT, in front of the different attention modules which are marked as Res-FCT+SENet, Res-FCT+ECA, Res-FCT+CAM, and Res-FCT+Res-CAM (HT-RCM) respectively.

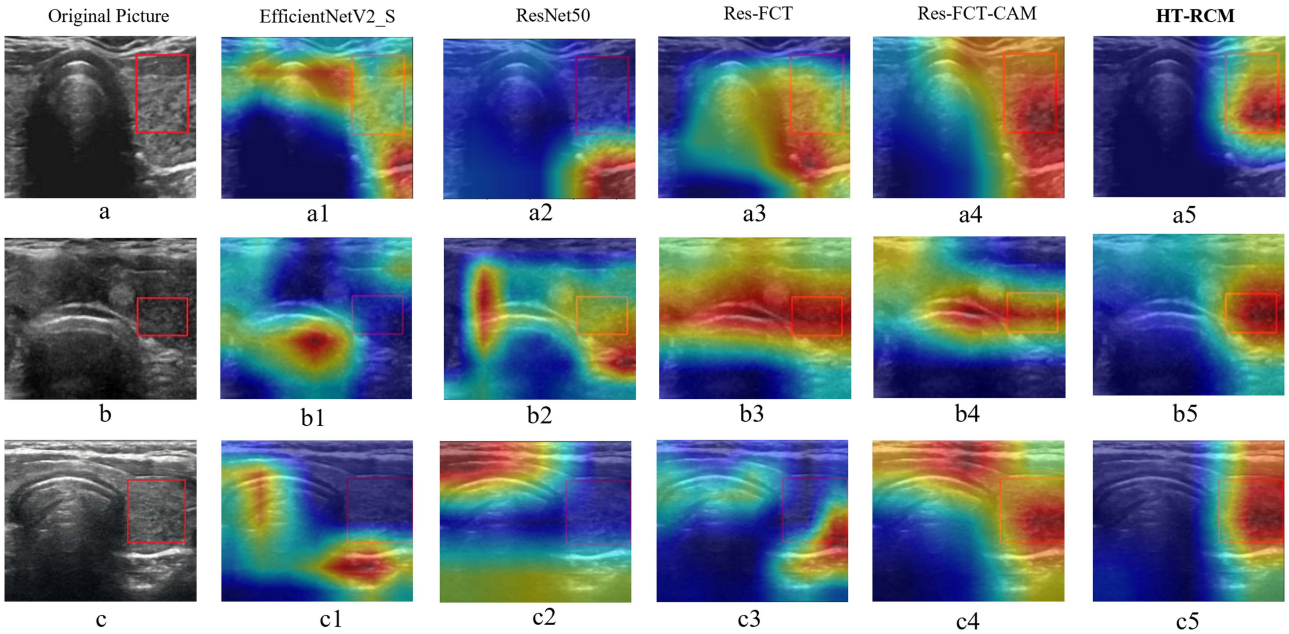

Fig. 5. Grad-CAM visualization results. We compare the visualization results of EfficientNetV2 S, ResNet50, Res-FCT, Res-FCT-CAM (Res-FCT + CAM), and HT-RCM (Res-FCT + Res-CAM). The grad-CAM visualization is calculated with the last convolutional outputs of each model.

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

TABLE III
EVALUATION OF HT-RCM WITH DIFFERENT NOISE

Model	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-score(%)	AUC
Data + Gauss noise	83.38	82.87	82.17	82.52	82.69	0.90
Data + White Gauss noise	83.80	83.15	81.23	82.60	82.87	0.90
Data + Salt and pepper noise	83.05	82.80	82.60	82.85	82.66	0.91

the thermal diagram was obtained using different models. As shown in Fig. 5(a1), the EfficientNetV2_S model learns image feature information in a wide range while covering a part of the noise region, and lacks central attentions to the lesion region. In Fig. 5(a2), the ResNet50 model focuses on learning part of the noise region while does not focus on the lesion region at all. In Fig. 5(a3), Res-FCT model effectively captures the global semantic information and long-distance semantic information and focuses on the lesion region accurately. However, it still focuses on some noise region. In Fig. 5(a4), the Res-FCT-CAM model further focuses on learning the features of the lesion region than Res-FCT model. However, there is still partial noise region is covered. In Fig. 5(a5), the HT-RCM focuses more on lesion features, reduces attention to the noise region, and realizes the accurate focusing on the lesion region. In Fig. 5(b1), the EfficientNetV2_S model does not learn the features of the lesion region completely while focuses on the noise region. In Fig. 5(b2), the ResNet50 model learns a wide range of image information, but misses learning the lesion region almost. In Fig. 5(b3), the Res-FCT model fully learns the image texture information while focuses too much on the noise region at the same time. In Fig. 5(b4), the Res-FCT-CAM model learns some features of the lesion region but pays also more attention to the noise region. In Fig. 5(b5), the HT-RCM fully extracts the features of the lesion region and effectively reduces the impacting of the noise region. In Fig. 5(c1), the EfficientNetV2_S model focuses on noisy regions and does not learn features of lesion regions completely. In Fig. 5(c2), the focus regions of the ResNet50 model are all noise regions. In Fig. 5(c3), the Res-FCT model focuses on a small part of the lesion region and a large part of the noise region, and so can not exactually extract the features of the lesion region. In Fig. 5(c4), the Res-FCT-CAM model can focus on the lesion region while cover large amount of noise region at the same time. Fig. 5(c5) shows that the brightness region cover and only cover the lesion region accurately which indicates that the HT-RCM can effectively exact and learn the characteristics of the lesion region. We can conclude through Fig. 5 that HT-RCM obtain the best focusing ability to the lesion region for HT ultrasound image classification.

G. Evaluation of HT-RCM With Different Noise

In order to further validate the performance of HT-RCM with different noisy levels, we add Gaussian noise, Gaussian white noise, and pepper and salt noise to the initial dataset respectively, and conduct experiments based on these dataset with different noise. The experimental results are shown in Table III. From Table III, we can find that all metrics of HT-RCM remain relatively stable, which proves that HT-RCM can still maintain good performance with different quality of Hashimoto thyroiditis ultrasonic image.

H. Results and Analysis

In order to analyze and compare the classification performance of EfficientNetV2_S, ResNet50, Res-FCT, Res-FCT-CAM, and HT-RCM models on HT ultrasound image, multiple metrics as given in (13)–(16) are used for the model evaluation. In the experiments, each model is trained for 10 rounds, with each round of 250 epochs, and the average accuracy is obtained when the model converges. Then, we train the HT-RCM and compare it with other methods on our dataset from four branch hospitals of Guangdong Hospital of Traditional Chinese Medicine (TCM). The experimental results are shown in Table IV. We can see from Table IV that the Accuracy, Sensitivity, Precision, F1-score, AUC, Params and Flops of HT-RCM are superior to the other methods. Specificially, the Accuracy, Sensitivity, Specificity, Precision, F1-score and AUC of HT-RCM are 1.65%, 2.43%, 0.92%, 0.60%, 1.51% and 0.03 higher than that of HTC-Net respectively, while Params and Flops are 15.44 M and 3.28 G lower than HTC-Net respectively, proving that HT-RCM has good generalization performance.

V. DISCUSSION

Hashimoto's Thyroiditis (HT) is a diffuse thyroid disease, and clinical researches have shown an increasing trend in the number of HT patients in recent years. The early and mid-term lesions of HT are not clearly characterized under ultrasound examination, and are easily confused with nodules, especially benign nodules. Benign thyroid nodules generally do not require special treatments, but HT patients must be reminded to keep better lifestyle habits and undergo regular examinations. Therefore, once a misdiagnose happens, it will delay the necessary treatments. As the condition worsens, it is easy to develop into more serious diseases such as hyperthyroidism, hypothyroidism, and even myxedema. The diffuse characteristics of early HT and the unclear imaging characteristics causing by hypoechoic signals pose serious challenges to the ultrasound diagnosis of HT. To accurately diagnose early HT through ultrasound technology, not only the low-order fine features but also the high-order distribution features of the ultrasound image need to be considered.

However, traditional CNNs or transformer models are either effective in extracting local information or in capturing global information, which exposes weakness in accurate clinical diagnosis of HT through ultrasound image. This paper proposes an HT ultrasound image classification model, named HT-RCM, based on Res-FCT and Res-CAM. HT-RCM aims to locate accurately the lesion region in the ultrasound image, extract local

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON OF EXPERIMENTAL RESULTS FROM DIFFERENT MODELS

Model	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-score(%)	AUC	Params(M)	Flops(G)
DenseNet	79.75	79.51	80.23	79.51	79.51	0.87	7.98	2.90
AlexNet	80.07	79.64	80.71	80.12	79.88	0.87	58.29	1.13
EfficientNetV2_S	80.48	82.56	78.46	78.70	80.58	0.87	21.46	2.90
ResNet50	82.06	83.35	81.71	79.64	81.45	0.88	25.56	4.13
ResNet152	77.27	75.59	79.19	79.51	77.50	0.86	60.19	11.60
HTC-Net	82.43	82.20	82.65	81.96	82.08	0.89	21.29	3.68
Res-FCT	82.73	82.71	82.74	81.95	82.34	0.90	**5.84**	0.40
Res-FCT-CAM	83.33	83.96	82.75	81.65	82.79	0.90	5.85	0.40
HT-RCM	**84.08**	**84.63**	**83.57**	**83.56**	**83.59**	**0.92**	5.85	**0.40**

The bold sections indicate that the individual models that performed best in the experiment are highlighted in bold.

fine features and global distribution features of the lesions, and enhance classification accuracy and generalization ability. The main components of HT-RCM are Res-FCT and Res-CAM.

To extract the low-order information produced by hypoechoic signals in ultrasound images, the Res-FCT model is designed by injecting the residual connection between Fully Convolutional Transformer (FCT). The local fine-grained lesion features mainly manifest as the low-order information, and the global lesion distribution features mainly manifest as the high-order information in the ultrasound images. Superimposing the low-order input information and high-order output information effectively can help HT-RCM obtain better feature representation while avoiding diffusion of gradients. In addition, to make HT-RCM learn multi-level original semantic information, a down-sampling module conducted on the initial feature map is injected into the corresponding FCT model. To help HT-RCM focus on the lesion regions more accurately, the Res-CAM is designed by implementing residual connection between the channel attention module (CAM) and the convolution layer. The Res-CAM can increase the weights of the lesion channels while suppressing the weights of the noise channels through adapting the channel weights dynamically. In addition, through enhancing the characteristics of the lesion region which is not evident and suppressing the noise region which is noticeable, Res-CAM can help HT-RCM improve classification accuracy and generalization ability. Extensive experiments demonstrate that our model achieves state-of-the-art performance in HT ultrasound image classification tasks compared with other mainstream models.

Although our model exhibits higher performance compared other mainstream models, the average accuracy is only 84.08% and the AUC is 0.92 which is still not high enough. The main reasons include: 1) The hypoechoic signal features of ultrasound images are not obvious which leads to the blurred imaging characteristics; 2) The early lesions of HT are very subtle and difficult to be detected; 3) The noise region is too large which surrounds the actual lesion region and affects the accuracy of the feature extraction methods; 4) The anti-interference ability of the model is not high enough which cannot effectively distinguish the HT lesions from other interfering lesions. So, future works can be carried out from the following directions to improve the proposed model. First, we shall improve the feature representation ability of hypoechoic signals from the perspective of ultrasound imaging technology and enhance the imaging features of hypoechoic signals. Second, due to the diffusing characteristics of HT lesions which mainly manifest as a large number of small lesions distributed globally. We can achieve feature filtering from the perspective of information filtering, attempting to filter out other interfering signals. Third, replacing the non-HT lesion regions with other regions in the same ultrasound image to avoid the impact of other lesions. Fourth, according to the ultrasound data characteristics, optimizing the model and its loss function, designing and realizing a more efficient neural network model.

VI. CONCLUSION

The early lesions of Hashimoto's thyroiditis (HT) are not evident, and the ultrasound image of HT are complex with large area of noises. This paper proposes a novel HT ultrasound image classification model HT-RCM based on Res-FCT and Res-CAM. Res-FCT can enhance the model's ability to extract both the long-distance and the global semantic information, and the residual connections injected in Res-FCT can helop to superimpose the low-order input information and high-order output information, reducing the lose of the original hypoechoic features. In addition, the initial feature map is down-sampled and reintegrated into the Res-FCT to further improve the learning ability to extract the original hypoechoic information. The Res-CAM is designed and constructed to help the model focus more attentions on the lesion region with inapparent features while reducing the attentions to the noise region. The experiments are implemented on 8000 ultrasound images collected by the endocrinology departments of four branch hospitals in Guangdong Provincial Hospital of Chinese Medicine. The experimental results show that HT-RCM exhibits the best overall performance compared with the current mainstream models. Furthermore, the training and learning process of HT-RCM does not need large-scale and well-annotated dataset, and the dataset used in this paper is raw data which only own 1 (HT) or 0 (no-HT) labels without any other annotation or lesion segmentation. So, it has high application and promotion value in clinical diagnosis.

VII. CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] M. I. A. Almalali et al., "Autoimmune thyroiditis a fibrous variant of Hashimoto's thyroiditis: A rare case," *Ann. Med. Surg.*, vol. 79, 2022, Art. no. 104019.

[2] N. R. Bayomy, M. A. Shaaban, A. E.-D. A. S. Dawood, M. E. A. Habib, and M. A. Kamel, "Correlation between circulating MicroRNA-142-5p expression and Hashimoto's thyroiditis diagnosis and autoimmunity symptoms prediction, pilot study," *Gene Rep.*, vol. 26, 2022, Art. no. 101470.

[3] H. Guan et al., "Discordance of serological and sonographic markers for Hashimoto's thyroiditis with gold standard histopathology," *Eur. J. Endocrinol.*, vol. 181, pp. 539–544, 2019.

[4] D. Wang et al., "Evaluation of thyroid nodules with coexistent Hashimoto's thyroiditis according to various ultrasound-based risk stratification systems a retrospective research," *Eur. J. Radiol.*, vol. 131, 2020, Art. no. 109059.

[5] X. Shen, C. Jiang, L. Sakhanenko, and Q. Lu, "A sieve quasi-likelihood ratio test for neural networks with applications to genetic association studies," 2022, *arXiv:2212.08255*.

[6] S. Qiao et al., "A pseudo-Siamese feature fusion generative adversarial network for synthesizing high-quality fetal four-chamber views," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 3, pp. 1193–1204, Mar. 2023.

[7] K. Wright et al., "The clinical significance of the American College of Radiology (ACR) thyroid imaging reporting and data system (TI-RADS) category 5 thyroid nodules: Not as risky as we think?," *Surgery*, vol. 173, no. 1, pp. 239–245, 2023.

[8] E. C. Bitkin and N. Toprak, "Diagnostic role of thyroid elastography in pediatric patients with Hashimoto's thyroiditis and graves' disease: A prospective controlled study," *Arch. de Pédiatrie*, vol. 30, no. 2, pp. 104–108, 2023.

[9] L. Anderson et al., "Hashimoto thyroiditis: Part 1, sonographic analysis of the nodular form of Hashimoto thyroiditis," *Amer. J. Roentgenology*, vol. 195, no. 1, pp. 208–215, 2010.

[10] X. Hu, Y. Chen, Y. Shen, R. Tian, Y. Sheng, and H. Que, "Global prevalence and epidemiological trends of Hashimoto's thyroiditis in adults: A systematic review and meta-analysis," *Front. Public Health*, vol. 10, 2022, Art. no. 1020709.

[11] M. Ralli et al., "Hashimoto's thyroiditis: An update on pathogenic mechanisms, diagnostic protocols, therapeutic strategies, and potential malignant transformation," *Autoimmunity Rev.*, vol. 19, no. 10, 2020, Art. no. 102649.

[12] X. Zhang, V. C. Lee, J. Rong, J. C. Lee, and F. Liu, "Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography," *Comput. Methods Programs Biomed.*, vol. 220, 2022, Art. no. 106823.

[13] Y. Wang et al., "TANet: A new paradigm for global face super-resolution via transformer-CNN aggregation network," 2021, *arXiv:2109.08174*.

[14] E. Z. Chen, C Zhang, X. Chen, Y. Liu, T. Chen, and S. Sun, "Computationally Efficient 3D MRI Reconstruction with Adaptive MLP," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258987249

[15] T. Tong et al., "Dual-input transformer: An end-to-end model for preoperative assessment of pathological complete response to neoadjuvant chemotherapy in breast cancer ultrasonography," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 251–262, Jan. 2023.

[16] J.-W. Hsieh, C.-H. Lee, Y.-C. Chen, W.-S. Lee, and H.-F. Chiang, "Stage classification in chronic kidney disease by ultrasound image," in *Proc. 29th Int. Conf. Image Vis. Comput. New Zealand*, 2014, pp. 271–276.

[17] M. Shi, J. Ding, S. Zhao, and Z. Huang, "Automatic thyroid ultrasound image detection and classification with priori knowledge," in *Proc. 5th Int. Conf. Comput. Sci. Appl. Eng.*, 2021, pp. 1–6.

[18] M. Song and Y. Kim, "Dual hybridization method for the classification of ultrasound breast tumors," in *Proc. 37th ACM/SIGAPP Symp. Appl. Comput.*, 2022, pp. 979–986.

[19] Y. Zhao, I. Tobore, W. Yan, and D. Que, "A computer-aided diagnosis system of breast lesion classification based on multi angle fusion strategy in ultrasound images," in *Proc. 2nd Int. Conf. Video, Signal Image Process.*, 2021, pp. 42–47.

[20] X. Li et al., "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study," *Lancet Oncol.*, vol. 20, no. 2, pp. 193–201, 2019.

[21] K. Gao et al., "Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101836.

[22] T. Zhang et al., "MSHT: Multi-stage hybrid transformer for the ROSE image analysis of pancreatic cancer," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1946–1957, Apr. 2023.

[23] X. Sheng, J. Chen, Y. Liu, B. Hu, and H. Cai, "Deep manifold harmonic network with dual attention for brain disorder classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 131–142, Jan. 2023.

[24] U. Acharya et al., "Diagnosis of Hashimoto's thyroiditis in ultrasound using tissue characterization and pixel classification," *Proc. Inst. Mech. Engineers, Part H: J. Eng. Med.*, vol. 227, no. 7, pp. 788–798, 2013.

[25] U. R. Acharya et al., "Computer-aided diagnostic system for detection of Hashimoto thyroiditis on ultrasound images from a polish population," *J. Ultrasound Med.*, vol. 33, no. 2, pp. 245–253, 2014.

[26] Q. Zhang et al., "Deep learning to diagnose Hashimoto's thyroiditis from sonographic images," *Nature Commun.*, vol. 13, 2022, Art. no. 3759.

[27] W. Zhao, Q. Kang, F. Qian, K. Li, J. Zhu, and B. Ma, "Convolutional neural network-based computer-assisted diagnosis of Hashimoto's thyroiditis on ultrasound," *J. Clin. Endocrinol. Metab.*, vol. 107, pp. 953–963, 2022.

[28] Z. Liang et al., "HTC-Net: Hashimoto's thyroiditis ultrasound image classification model based on residual network reinforced by channel attention mechanism," *Health Inf. Sci. Syst.*, vol. 11, 2023, Art. no. 24.

[29] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, 2022, Art. no. 102357.

[30] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, "The fully convolutional transformer for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3649–3658.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[32] E. M. Bakr, A. El-Sallab, and M. Rashwan, "EMCA: Efficient multiscale channel attention module," *IEEE Access*, vol. 10, pp. 103447–103461, 2022.

[33] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6304–6308.

[34] J. Zhuang, Y. Dong, H. Bai, P. Zuo, and J. Cheng, "Auto-selecting receptive field network for visual tracking," *IEEE Access*, vol. 7, pp. 157449–157458, 2019.

[35] Y. Deng, X. Li, M. Zhang, X. Lu, and X. Sun, "Enhanced distance-aware self-attention and multi-level match for sentence semantic matching," *Neurocomputing*, vol. 501, pp. 174–187, 2022.

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

[37] J. Zhou, Y. Du, R. Zhang, and R. Zhang, "Adaptive depth graph attention networks," vol. abs/2301.06265, 2023, *arXiv:2301.06265*.

[38] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[39] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[41] M. K. Jalehi and B. M. Albaker, "Highly accurate multiclass classification of respiratory system diseases from chest radiography images using deep transfer learning technique," *Biomed. Signal Process. Control*, vol. 84, 2023, Art. no. 104745.

[42] M.-L. Huang and Y.-C. Liao, "Stacking ensemble and ECA-EfficientNetV2 convolutional neural networks on classification of multiple chest diseases including COVID-19," *Academic Radiol.*, vol. 30, no. 9, pp. 1915–1935, 2023.

[43] A. Panthakkan, S. Anzar, S. Jamal, and W. Mansoor, "Concatenated Xception-ResNet50—A novel hybrid approach for accurate skin cancer prediction," *Comput. Biol. Med.*, vol. 150, 2022, Art. no. 106170.

[44] N. N. Prakash, V. Rajesh, D. L. Namakhwa, S. D. Pande, and S. H. Ahammad, "A densenet CNN-based liver lesion prediction and classification for future medical diagnosis," *Sci. Afr.*, vol. 20, 2023, Art. no. e01629.

[45] H. Eldem, E. Ülker, and O. Y. Işıklı, "AlexNet architecture variations with transfer learning for classification of wound images," *Eng. Sci. Technol., Int. J.*, vol. 45, 2023, Art. no. 101490.

[46] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, pp. 336–359, 2016.

Authorized licensed use limited to: Universiteit van Amsterdam. Downloaded on June 17,2024 at 13:22:20 UTC from IEEE Xplore. Restrictions apply.