

# Hierarchical Temporal Attention Network for Thyroid Nodule Recognition Using Dynamic CEUS Imaging

Peng Wan<sup>1</sup>, Fang Chen<sup>1</sup>, Chunrui Liu, Wentao Kong, and Daoqiang Zhang<sup>1</sup>, *Member, IEEE*

**Abstract**—Contrast-enhanced ultrasound (CEUS) has emerged as a popular imaging modality in thyroid nodule diagnosis due to its ability to visualize vascular distribution in real time. Recently, a number of learning-based methods are dedicated to mine pathological-related enhancement dynamics and make prediction at one step, ignoring a native diagnostic dependency. In clinics, the differentiation of benign or malignant nodules always precedes the recognition of pathological types. In this paper, we propose a novel hierarchical temporal attention network (HiTAN) for thyroid nodule diagnosis using dynamic CEUS imaging, which unifies dynamic enhancement feature learning and hierarchical nodules classification into a deep framework. Specifically, this method decomposes the diagnosis of nodules into an ordered two-stage classification task, where diagnostic dependency is modeled by Gated Recurrent Units (GRUs). Besides, we design a local-to-global temporal aggregation (LGTA) operator to perform a comprehensive temporal fusion along the hierarchical prediction path. Particularly, local temporal information is defined as typical enhancement patterns identified with the guidance of perfusion representation learned from the differentiation level. Then, we leverage an attention mechanism to embed global enhancement dynamics into each identified salient pattern. In this study, we evaluate the proposed HiTAN method on the collected CEUS dataset of thyroid nodules. Extensive experimental results validate the efficacy of dynamic patterns learning, fusion and hierarchical diagnosis mechanism.

**Index Terms**—Contrast-enhanced ultrasound, hierarchical, temporal attention, thyroid nodule.

## I. INTRODUCTION

CONTRAST enhanced ultrasound (CEUS) has been one of the most advanced imaging techniques in clinical tumor treatments, ranging from early screening [1], differential diagnosis [2] to treatment response evaluation [3]. In comparison to conventional B-mode ultrasound (BUS), CEUS is particularly useful for the detection and characterization of lesions, due to its ability to highlight blood flow in microvasculature [4]. Through the use of gas-filled microbubble contrast agents (CAs) [5], [6], CEUS enables radiologists to monitor the relative echoic intensity changes<sup>1</sup> between lesion regions and the surrounding tissues in real-time [7]. The dynamic enhancement process from wash-in to wash-out is shown in Fig. 1(a). Due to its close correlation with intra-tumor neoplastic cell growth [7], [8], dynamic enhancement patterns contained in CEUS sequences form the diagnostic basis. Taking thyroid nodules as example, hypo-enhancement and heterogeneous enhancement are considered to be two major indicators of malignancy, especially for nodules with a diameter of 10 mm or less, while homogeneous and ring enhancement are two typical patterns for benign ones [9]. However, current clinical decision is heavily dependent on the expertise of the radiologists, due to the operator-dependent diagnostic procedure (e.g., identifying an optimal frame where the lesion boundary is well-distinguished from surrounding parenchyma; assessing perfusion patterns by observing the whole sequence back and forth). Thus, various computer-aided diagnosis (CAD) approaches have been proposed to improve diagnostic performance and ease the clinical workload with the explosion of CEUS data [10]–[12].

Compared with conventional static BUS imaging analysis, automatic interpretation of dynamic CEUS imaging series could be more challenging, as rapidly changing contrast frames often show diverse enhancement patterns (e.g., inhomogeneous enhancement, rim enhancement, spoke-wheel enhancement, etc.) at different perfusion stages. Furthermore, for different

Manuscript received January 25, 2021; revised February 21, 2021; accepted February 28, 2021. Date of publication March 2, 2021; date of current version June 1, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61876082, Grant 61732006, Grant 61861130366, Grant 61901214, Grant U20A20389, and Grant 81671701; in part by the National Key Research and Development Program of China under Grant 2018YFC2001600, Grant 2018YFC2001602, and Grant 2018ZX10201002; in part by the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship under Grant NAF\R1\180371; in part by the China Postdoctoral Science Foundation funded project under Grant 2020M671484; and in part by the Nanjing Medical Science and Technique Development Foundation under Grant YKK19054. (Corresponding authors: Wentao Kong; Daoqiang Zhang.)

Peng Wan, Fang Chen, and Daoqiang Zhang are with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Chunrui Liu and Wentao Kong are with the Department of Ultrasound, Affiliated Drum Tower Hospital, Nanjing University Medical School, Nanjing 210008, China (e-mail: breezewen@163.com).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TMI.2021.3063421>, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3063421

<sup>1</sup>That is termed as dynamic enhancement patterns in the field of CEUS imaging analysis

types of lesions, the degree of enhancements and the speed of wash-in and wash-out are often numerous. In order to distill informative perfusion characteristics from CEUS videos, early researches extracted quantitative functional parameters from either regional time-intensity curves (TICs) [13] or factor curves obtained by matrix factorization [14], [15]. Nonetheless, these parameters can only describe limited functional information while abundant morphological features regarding tumor vascularization have been discarded. Subsequently, predefined high-order texture features are used to quantify the enhancement patterns by the multi-view learning framework [16], [17]. Further, three-dimensional convolutional neural network (3D-CNN) is applied to learn the task-oriented spatial-temporal representations in the data-driven manner [11], considering that handcrafted features might not be well-coordinated with subsequent classifier construction. We could observe that most existing studies have been dedicated to quantify temporal dynamics and make diagnostic prediction at one stage, either benign (malignant) differentiation or pathological recognition.

In fact, there exists a hierarchical structure for lesion diagnosis at different stages. As shown in Fig. 1(b), a hierarchical (coarse-to-fine) identification process is indispensable in clinical workflow [18]. Generally, radiologists first determine whether the lesion is benign or malignant according to several typical enhancement patterns; after that, they would further distinguish specific pathological types from each other within the candidate set of benign (malignant) types. Similar hierarchical prediction is first applied in caption generation [19] and vehicle re-identification [20], where sequential predictions are made conditioned on the previous context. Nonetheless, this hierarchical dependency relationship is yet to be taken into account in CEUS based diagnostic models. Another major limitation of these work is that each contrast frame is treated equally important for feature extraction and classifier construction, while inherent differences in the contributions of different frames are ignored. As known from clinical practices, radiologists usually observe the whole perfusion sequence back and forth at first, then focus on partial salient enhancement patterns according to their experiences. That is, distinctive temporal points should have different contributions with respect to the specific tumor type. To model this observation process, an effective temporal aggregation strategy that integrates local salient patterns and global dynamic evolutions should also be considered.

In this paper, we propose a *hierarchical temporal attention network (HiTAN)* method for identifying different types of nodules using dynamic CEUS imaging. To model the diagnostic dependency of different stages, the proposed HiTAN constructs a hierarchical classifier in which fined-grained nodule recognition is made conditioned on the knowledge learned from coarse-grained classification. Besides, the hierarchical relationship also consists in automatically identified local key frames, since the generation of temporal attention score is guided by the global perfusion representation extracted at the coarse-grained stage. As shown in Fig. 2, our proposed HiTAN consists of three fundamental components, 1) Frame-level enhancement representation learning module; 2) Hierarchical

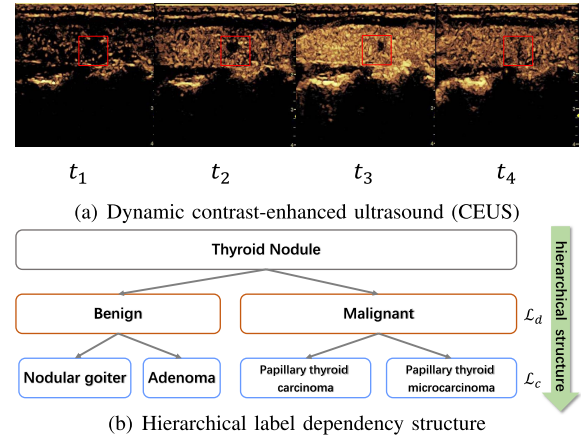


Fig. 1. Fig 1(a) shows the sampled dynamic contrast-enhanced ultrasound (CEUS) frames from a malignant thyroid nodule. The relative intensity variation differences between lesions (outlined in red boxes) and surrounding healthy tissues form the diagnostic basis. Fig 1(b) presents the hierarchical label dependency relationship of thyroid nodules recognition. In clinical workflow, radiologists usually first perform the differentiation of benign or malignant nodules  $L_d$ ; Based on that, they would further distinguish specific pathological type from each other within the candidate set of benign (malignant) types  $L_c$ .

lesion recognition module; 3) Local-Global temporal aggregation module. Following the description in [21], we term the coarse-grained identification of benign (malignant) as “differentiation”, and the fined-grained pathological recognition as “characterization”. We evaluate our method on our collected datasets of thyroid nodules. Experimental results demonstrated that our proposed HiTAN method not only can achieve superior diagnostic performance compared with those competing methods, but also can identify salient perfusion patterns of great diagnostic value.

## II. RELATED WORK

In this section, we briefly review two lines of previous work on CEUS-based automatic tumor diagnosis, including the conventional parameters-based methods and learning-based methods.

### A. Parameters-Based Methods

Parameters-based methods are the most common techniques in current clinical applications, which depend on multiple kinetic parameters extraction (e.g., Peak value, Time-to-peak and Area under curve, etc.) from regional time intensity curves (TICs). Numerous clinical studies find that there exist significant statistical correlations between tumor microvascularization and TIC parameters [22], [23]. For example, Huang-Wei *et al.* [24] extracted TIC from manually outlined regions of interest (ROIs) and confirmed that regional TIC parameters could help differentiate focal nodular hyperplasia (FNH) from other tumors such as hepatocellular carcinoma (HCC), liver metastasis, and hepatic hemangioma by statistical analysis. Other regional TIC analysis methods also attempted to train basic models (e.g., support vector machine (SVM), k-nearest neighbors (KNN) and artificial neural network (ANN)) using limited kinetic features for

lesion classification [10], [25]. Some researches stated that simply averaging regional intensities inevitably loses the intra-tumor heterogeneity information. Factor analysis methods (e.g. sparse non-negative matrix factorization (Sparse NMF)) have also been used to distill representative physiological curves from a number of pixel-wise curves [14], [26]. The major limitation of parameters-based pattern analysis is that such description manner relying on few curve kinetic parameters is too simple to encode abundant texture features in different enhancement stages.

### B. Learning-Based Methods

In order to enhance the representation power for diverse enhancement patterns, a growing number of learning-based methods resort to more abundant texture features or data-driven convolution features. For example, Guo *et al.* [17] employed the multi-view learning framework to fuse enhancement patterns at different temporal phases, in which each temporal segment was treated as one view. Shared latent representations learned from each pair of phases were fed into a multiple kernel learning (MKL) classifier for the final prediction. In [16], the static B-mode view was also incorporated for further texture feature fusion and DCCA (a deep variant of CCA) was introduced to improve nonlinear representation power. However, these works still depended on the manual selection of representative contrast frame at different phases, which was time-consuming and prone to subjective errors. To achieve an automatic spatial-temporal pattern search and fusion, Liang *et al.* [21] proposed a novel latent structured model which could selectively combine considerable candidate ROI spread over the whole perfusion sequence. Notably, to cohere with the clinical observation, region representation in [21] was constructed based on texture features both in the interior and surrounding of ROIs.

Nonetheless, hand-crafted features are independent of subsequent dynamic feature fusion, thus may not be well coordinated with subsequent classifier construction, leading to a sub-optimal diagnostic performance. Inspired by the powerful visual representation capacity of deep learning, recent studies began to utilize deep convolution neural network (DCNN) to extract task-oriented, high-nonlinear features for lesion recognition. Given dynamic enhancement patterns in CEUS videos, temporal modeling is the core part of deep learning-based methods. One of the most direct way is to extend 2D CNNs to their 3D counterparts [27], where a series of 3D convolution filters are stacked to jointly model spatial and temporal enhancement characteristics reflecting angiogenesis [11]. To avoid over-fitting risk caused by considerably increased parameters, several variants of 3D CNN (i.e., I3D [28] and R2Plus1D [29]) are proposed, which utilize pre-trained 2D CNN to learn frame-level features, followed by a temporal aggregation module for post-hoc fusion. 1D pooling [30] and recurrent neural networks (RNN) [31], [32] are commonly-used techniques to obtain video-level representations. Among which, attention mechanism is used to attend to those salient frame-level representations [33], [34]. To establish finer-grained interactions between frames even distant in time,

a number of temporal dependency modeling strategies are proposed in recent years, including non-local mechanism [35], temporal relation reasoning [36], space-time region graphs [37], long-term feature bank [38], and temporal shift module [39], etc. Inspired by 3D CNN where temporal information is aggregated at various visual tempos at different depths, another line of work attempt to construct temporal pyramid network to handle the variance visual tempos [40]–[42]. Although both conventional and deep learning based methods aimed at better dynamic perfusion characteristics representation, most of them ignore a basic diagnostic dependency relationship (i.e., label hierarchical structure) in their model design. In fact, such hierarchical structure of diagnostic tasks can be utilized to design an effective guidance mechanism in dynamics learning.

## III. METHOD

In this part, we introduce the proposed hierarchical temporal attention network (HiTAN) in detail, including the overall representation learning and hierarchical recognition architecture (Section III-A), a specific loss for joint differentiation and characterization (Section III-B), and the implementation details (Section III-C).

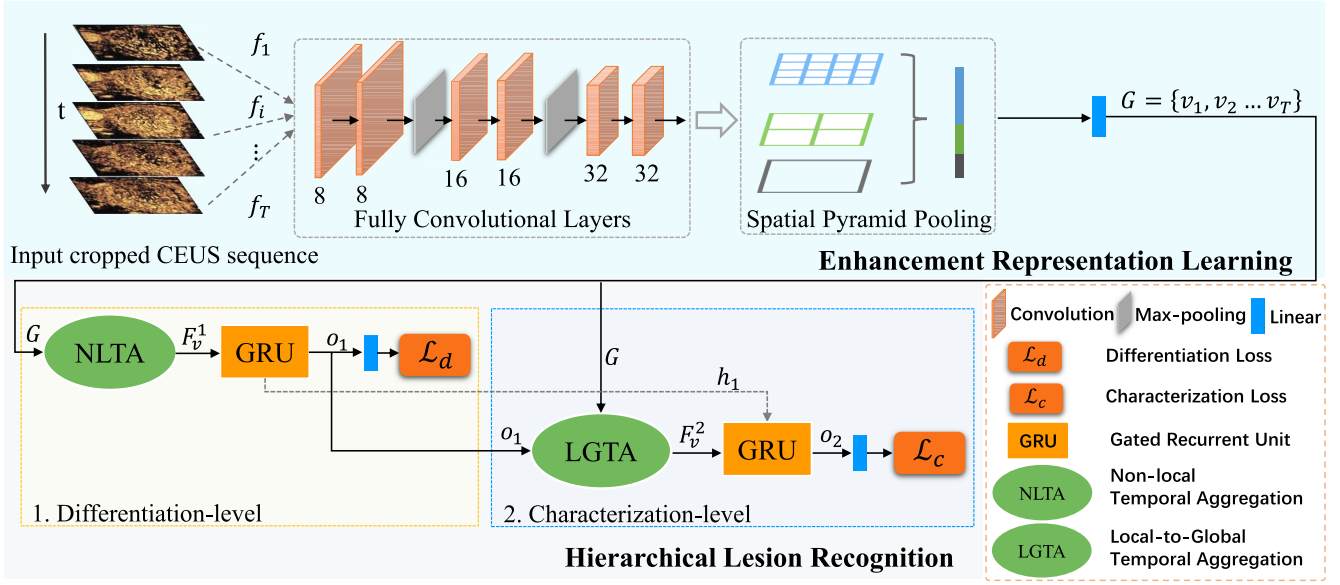
### A. Architecture

As shown in Fig. 2, our proposed hierarchical temporal attention network (HiTAN) consists of two main components, including 1) Enhancement representation learning module and 2) Hierarchical lesion recognition module.

Briefly, the sequential CEUS frames are *independently* forwarded through the enhancement representation learning module to produce enhancement descriptor  $v_t$  at each time step; Then, the coarse-to-fine diagnostic procedure is modeled by Gated Recurrent Units (GRUs) which allows the pathological prediction  $\mathcal{L}_c$  to be made conditioned on the differentiation of benign and malignant  $\mathcal{L}_d$ . To integrate dynamic enhancement patterns throughout the perfusion process, *at the differentiation level*, sequential enhancement descriptors  $\mathbf{G}$  are aggregated with the non-local temporal aggregation (NLTA) operator, which considers dense inter-frame relations between all pairs of points. *At the characterization level*, we propose a novel local-global temporal aggregation (LGTA) operator to realize a comprehensive analysis of local representative patterns and global enhancement dynamics. The architecture of our proposed HiTAN is detailed as follows.

**1) Enhancement Representation Learning:** In order to extract effective temporal dynamics, the first step is to locate informative perfusion regions. To this end, we adopt the CEUS-Net from our previous work [43] to perform lesion segmentation on dynamic CEUS sequences. To alleviate distribution discrepancy, we choose to fine-tune the network parameters of CEUS-Net (Detailed in Fig. S1 and Section A in Supplementary Materials), thus reducing the number of required annotated samples. Then, a segmentation mask is used to crop the sequential CEUS frames by generating a bounding box enclosing the foreground. Notably, pixel-wise annotation is labelled on the contrast frame around perfusion peak after motion correction (Detailed in Section III-C.1).





**Fig. 2.** Illustration of the framework of Hierarchical Temporal Attention Network (HiTAN), which consists of two fundamental modules, 1) Enhancement Representation Learning and 2) Hierarchical Lesion Recognition module. As presented, the hierarchical recognition module decomposes the diagnosis of thyroid nodules into two steps, i.e., from the coarse-grained benign (malignant) *differentiation* to the fine-grained pathological *characterization*. Particularly, the native diagnostic dependency is modeled by Gated Recurrent Units (GRUs), where internal hidden state  $h_1$  is expected to maintain enhancement dynamics learned from the differentiation level at step 1 and benefit the pathological characterization at step 2. To capture long-range temporal cues at different levels, we introduce the Non-local Temporal Aggregation (NLTA) and Local-to-Global Temporal Aggregation (LGTA) operators to aggregate the sequence of enhancement descriptors  $\mathbf{G}$ ,  $F_v^1 = \text{NLTA}(\mathbf{G}, \mathbf{G})$  and  $F_v^2 = \text{LGTA}(\mathbf{L}(o_1), \mathbf{G})$ , where the output of GRUs at the first step  $o_1$  also functions as the source to produce temporal guidance signal that assists the identification of salient subsequence  $\mathbf{L}$ .

According to clinical knowledge, contrast variations of surrounding tissues also provide important reference to that of lesion interior [10], [21], we enlarge the bounding box by a small factor  $\gamma = 1.2$  to incorporate part of neighboring tissues.  $\gamma$  is a hyperparameter that we cross-validate in experiments.

Specifically, fully convolutional layers are composed of six  $3 \times 3$  convolutional (Conv) layers and two  $2 \times 2$  local pooling layers. All Conv layers have unit stride with zero-padding, followed by the BatchNorm and ReLu activation. The number of channels for Conv1 to Conv6 are 8, 8, 16, 16, 32, and 32, respectively. Among which, Conv2 and Conv4 are followed by a max-pooling layer with stride 2 to down-sample the intermediate feature maps while increase the receptive fields. Notably, another benefit of local pooling is to further alleviate the impact of residual tissue motion in spite of motion compensation. Considering the varying sizes of cropped regions, the resulting convolutional maps are fed into a spatial pyramid pooling (SPP) layer [44], aiming at producing a fixed-length enhancement descriptor  $v_t$  via adaptive pooling of all channels. By adopting a four-level pyramid  $\{4 \times 4, 2 \times 2, 1 \times 1\}$ , we could obtain a  $21 \times 32$  dimensional embedding  $v_t$ . Finally, we use a linear layer to reduce the dimensionality to  $d = 128$ , followed by a dropout layer with rate 0.1 [41], [42], [45]. As a plug-in unit, this basic backbone in our HiTAN can be easily replaced by any other lightweight convolutional architectures allowing for varying-size inputs.

**2) Hierarchical Lesion Recognition:** In the hierarchical lesion recognition module, we decompose the diagnosis of thyroid nodules into an ordered two-stage classification, i.e., from the coarse-grained benign (malignant) *differentiation* to

the fine-grained pathological *characterization*. To model the native diagnostic dependency, we employ the Gated Recurrent Units (GRUs) to connect two consecutive classification tasks [19], where the inner hidden state  $h_1$  at the characterization level is expected to maintain valuable perfusion information learned from the differentiation level. In this way, we could make the pathological prediction conditioned on the differentiation of benign or malignant nodules. In parallel, since enhancement characteristics of different frames often show large variations across different perfusion phases, *how to aggregate diverse enhancement patterns over long temporal support* is another major challenge to this work.

**Non-local Temporal Aggregation:** At the differentiation level, we adopt the non-local temporal aggregation (NLTA) operator [35] to fuse dynamic enhancement descriptors  $\mathbf{G} = \{v_1, v_2, \dots, v_T\} \in R^{d \times T}$ , where  $T$  is the length of input CEUS sequences. As shown in Fig. 3(a), by inferring contextual relations between all pairs of points even distant in time, NLTA can provide long-term supportive enhancement information to  $t$ -th frame descriptor  $v_t$ . The NLTA operator can be formulated as follows,

$$c_{t,i} = \frac{\exp(v_t^T \mathbf{W}_1 v_i) + \varepsilon_1}{\sum_{i=1}^{i=T} \exp(v_t^T \mathbf{W}_1 v_i) + \varepsilon_1} \quad (1)$$

$$\tilde{v}_t = v_t + \mathbf{H}_1 \sum_{i=1}^{i=T} c_{t,i} \mathbf{P}_1 v_i \quad (2)$$

$$F_v^1 = \frac{1}{T} \sum_{t=1}^{t=T} \tilde{v}_t \quad (3)$$

where  $\mathbf{W}_1 = \mathbf{U}_1^T \mathbf{V}_1$ ,  $\mathbf{U}_1, \mathbf{V}_1, \mathbf{H}_1, \mathbf{P}_1 \in R^{d \times d}$  are weight parameters of linear layers, superscript  $T$  denotes the transpose

operation.  $\varepsilon_1$  is a small constant  $10^{-5}$  for numerical stability.  $\tilde{v}_t$  is the updated descriptor that embeds global enhancement characteristics based on semantic correlations  $c_{t,i}$ .  $F_v^1$  denotes the aggregated enhancement representation via average pooling. Due to the limited CEUS data, layer normalization [46] and dropout with the rate of 0.1 [45] are appended to each linear layer for reducing the overfitting risk.

**Local-to-Global Temporal Aggregation:** By recalling clinical CEUS imaging analysis, radiologists usually perform a comprehensive analysis of local and global enhancement characteristics. The first step is to identify local salient enhancement patterns throughout the CEUS sequence, which reflect interior vessel distribution or periphery invasion, such as rim-like hyper-enhancement, heterogeneous enhancement, and iso- or hypo- enhancement, etc.; The second step is to further review the whole CEUS video back and forth, which aims at capturing overall variation tendency correlating with changes of blood hemodynamics, such as diffuse or concentric enhancement, rapid or slow wash- in or out, etc. Inspired by this, we design a novel Local-to-Global Temporal Aggregation (LGTA) operator to integrate local and global enhancement features. It is expected to comprehensively capture perfusion characteristics regarding intra-nodular vascularity at the characterization level. In this operator, local temporal information is defined as salient enhancement patterns identified with the guidance from differentiation-level perfusion representation, while global temporal information refers to overall enhancement tendency encoded in the whole sequences. In comparison to temporal pooling in [47], inter-frame relation is estimated in LGTA and restricted between local salient patterns and the whole contrast sequence, which differs from non-local mechanism in [35] that infers correlations between all pairs of frames. Also, different from the long-term feature bank operator in [38] that provides global temporal information to each fixed-length sliding window, our LGTA only embeds global enhancement dynamics into every identified keyframe.

As shown in Fig. 3(b), LGTA takes the output of GRUs at the differentiation level  $o_1$  and the sequence of enhancement descriptors  $\mathbf{G}$  as inputs, outputs the video-level perfusion representation  $F_v^2$  which embeds global temporal cues into local salient patterns. At the first step, we leverage the output of GRUs  $o_1$  learned at the differentiation stage as the source to generate the attention guidance signal  $g'$ , which assists the definition of local temporal supports in LGTA that contain discriminative enhancement characteristics. Therefore, the generation of temporal attention vector  $A$  can be expressed as,

$$A = \text{softmax} \left( o_1^T \mathbf{F}^T \mathbf{G} \right) \quad (4)$$

where  $A = [a_1, a_2, \dots, a_T] \in R^{1 \times T}$ ,  $\mathbf{G} = \{v_1, v_2, \dots, v_T\} \in R^{d \times T}$ ,  $g' = \mathbf{F}o_1$ ,  $\mathbf{F} \in R^{d \times d}$  is the weight matrix of the linear layer applied on  $o_1$ . The  $\text{softmax}$  function is used to normalize the  $t$ -th frame temporal weight  $a_t$  into  $[0, 1]$ ,

$$\text{softmax}(a_t) = \frac{\exp(a_t) + \varepsilon_2}{\sum_{t=1}^T \exp(a_t) + \varepsilon_2} \quad (5)$$

where  $\exp(\cdot)$  denotes the exponential function and  $\varepsilon_2$  is a small constant  $10^{-5}$ . In this way, when  $\mathbf{F}$  is an identity matrix,

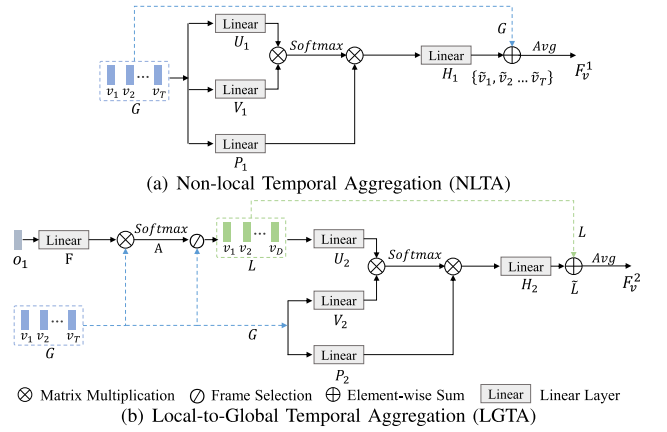


Fig. 3. Fig. 3(a) Non-local Temporal Aggregation (NLTA) operator: it considers all pairwise interactions between points even they are distant in time; Fig. 3(b) Local-to-Global Temporal Aggregation (LGTA) operator: it restricts inter-frame correlations between the identified local salient patterns  $\mathbf{L}$  and the whole sequence of enhancement descriptors  $\mathbf{G}$ . As can be seen, the output of LGTA is an updated enhancement representation  $\tilde{\mathbf{L}}$  that embeds global enhancement dynamics.

a particular frame descriptor  $v_t$  can bear contribution only in proportion to its compatibility with  $o_1$ , which is used for lesion differentiation at the differentiation level. That is, one CEUS frame should be attended only when it shows visually similar enhancement patterns encoded by  $o_1$ . Nonetheless, pathological recognition apparently requires finer-grained perfusion characteristics as the diagnostic basis. Therefore, a linear layer  $\mathbf{F}$  is used to transform  $o_1$  into a new latent space such that those different but equally important patterns could be attended to. Based on that, the former  $D$  frames with higher weights are selected as local temporal information  $\mathbf{L} = \{v_1, v_2, \dots, v_D\} \in R^{d \times D}$ .  $D$  is a hyperparameter determined by cross-validation in our experiments. In the current implementation,  $D$  is empirically set to 7.

The second step is to integrate local and global enhancement dynamics using an attention mechanism [35]. In short, we use local features  $\mathbf{L}$  to attend to sequential enhancement descriptors  $\mathbf{G}$ , and add attended global information back to  $\mathbf{L}$  via a residual connection, which can be summarized as follows,

$$\mathbf{C} = \text{softmax} \left( \mathbf{L}^T \mathbf{U}_2^T \mathbf{V}_2 \mathbf{G} \right)_{row} \quad (6)$$

$$\mathbf{Q} = \mathbf{L} + \mathbf{H}_2 \mathbf{P}_2 \mathbf{G} \mathbf{C}^T \quad (7)$$

$$F_v^2 = \frac{1}{D} \mathbf{Q} \mathbf{1} \quad (8)$$

where  $\mathbf{C} \in R^{D \times T}$  is the correlation matrix normalized row-by-row. In Eq. 7,  $\mathbf{C}$  is used to update local features  $\mathbf{L}$  by embedding global perfusion features provided by  $\mathbf{G}$ .  $\mathbf{U}_2, \mathbf{V}_2, \mathbf{H}_2, \mathbf{P}_2 \in R^{d \times d}$  are weight parameters of linear layers, and the superscript  $T$  denotes matrix transpose.  $\mathbf{1}$  is  $d$ -dimensional vector of ones,  $F_v^2$  is the aggregated temporal representation at the characterization level. Similarly, we add the layer normalization and dropout with the rate of 0.1 to liner layers to improve regularization.

**Coarse-to-fine nodule classification:** As shown in Fig. 2, we first aggregate successive enhancement descriptors  $\mathbf{G} = \{v_1, v_2, \dots, v_T\}$  using the NLTA operator. Then

the resulting video-level enhancement representation  $F_v^1 = NLTA(\mathbf{G}, \mathbf{G})$  is fed into GRUs, with the output  $o_1 = GRU(F_v^1, h_0)$  used to calculate the benign (malignant) score *at the differentiation level*. More importantly, *at the characterization level*, we also treat  $o_1$  as the source to generate temporal guidance signal, which assists to evaluate which part of the contrast frames encode finer-grained enhancement characteristics reflecting neovascularization of tumor growth. Based on the learned weights, we preserve a proportion of enhancement patterns as local temporal information  $\mathbf{L} = \{v_{1'}, v_{2'}, \dots, v_{D'}\}$ . After that, we further embed global enhancement characteristics into each identified local pattern based on contextual relations, thus the output  $F_v^2 = LGTA(\mathbf{L}(o_1), \mathbf{G})$  is considered as a local-to-global enhancement feature aggregation. Finally,  $F_v^2$  is fed into GRUs and the output  $o_2 = GRU(F_v^2, h_1)$  is used for pathological prediction. The inner update mechanism of GRUs at two levels  $l = \{1, 2\}$  can be formulated as:

$$z_l = \sigma(\mathbf{W}_z x_l + \mathbf{U}_z h_{l-1}) \quad (9)$$

$$r_l = \sigma(\mathbf{W}_r x_l + \mathbf{U}_r h_{l-1}) \quad (10)$$

$$o_l = \tanh(\mathbf{W}_o x_l + \mathbf{U}_o (r_l \circ h_{l-1})) \quad (11)$$

$$h_l = (1 - z_l) \circ h_{l-1} + z_l \circ o_l \quad (12)$$

where  $z_l$  and  $r_l$  denote the update gate and reset gate value, respectively, which are derived by the *sigmoid* activation function  $\sigma$ .  $\mathbf{W}$ ,  $\mathbf{W}_z$ ,  $\mathbf{W}_r$ ,  $\mathbf{U}$ ,  $\mathbf{U}_z$ , and  $\mathbf{U}_r \in \mathbb{R}^{d \times d}$  are learned parameters of linear transformations.  $o_l$  is the output activated by tangent function *tanh*. The symbol  $\circ$  represents the scalar product with a gate value. As known from Eq. (9 – 12), the inner gating mechanism enables GRUs to selectively preserve and update the hidden state  $h_l \in \mathbb{R}^d$ , revealing the knowledge transfer along the prediction path. In our implementation,  $h_0$  is initialized as a  $d$ -dimensional zero vector.

### B. Loss for Joint Differentiation and Characterization

Let  $\{(\mathbf{X}_n, \mathbf{y}_n)\}_{n=1}^N$  be the training set containing  $N$  samples, where  $\mathbf{X}_n = \{I_1, I_2, \dots, I_T\}$  and  $\mathbf{y}_n^c \in \{1, \dots, C\}$  denote the CEUS sequence for  $n^{th}$  subject and the corresponding pathological category ( $C = 4$ ), and  $\mathbf{y}_n^d \in \{1, 2\}$  is the benign(malignant) label. Our HiTAN model performs an attention-aware hierarchical lesion recognition based on dynamic CEUS sequences. It jointly optimizes the learnable parameters of the backbone network, the hierarchical recognition module and the local-global temporal aggregation module, respectively. As shown in Fig. 2, the loss functions  $\mathcal{L}$  is the sum of the loss for lesion benign (malignant) differentiation (i.e.,  $\mathcal{L}_d$ ) and the loss for pathological characterization (i.e.,  $\mathcal{L}_c$ ).

$$\begin{aligned} \mathcal{L} &= \lambda_d \mathcal{L}_d + \mathcal{L}_c \\ &= -\frac{\lambda_d}{N} \sum_{n=1}^N \sum_{d=1}^2 \mathbb{I}(\mathbf{y}_n^d = d) \log(\mathcal{P}^d(\mathbf{X}_n | \mathbf{W}^n, \mathbf{W}^d)) \\ &\quad - \frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C \mathbb{I}(\mathbf{y}_n^c = c) \log(\mathcal{P}^c(\mathbf{X}_n | \mathbf{W}^n, \mathbf{W}^c)) \end{aligned} \quad (13)$$

where  $\mathbb{I}(\cdot)$  is a binary indicator, the learnable parameters for the backbone network, differentiation-, and characterization-level subnetwork are denoted as  $\mathbf{W}^n$ ,  $\mathbf{W}^d$  and  $\mathbf{W}^c$  respectively.

$\mathcal{P}^c(\mathcal{P}^d)$  represents the score for classifying  $\mathbf{X}_n$  as the  $c$ -th category in terms of network parameters.  $\lambda_d$  is the trade-off parameter that balances the importance of two-stage predictions.  $\lambda_d$  is a hyperparameter that we cross-validated in our experiments, and empirically set to 0.25 in our current implementation.

### C. Implementations

The proposed HiTAN model was implemented using the popular deep learning framework Pytorch, and codes were run on the single GPU (i.e., NVIDIA TITAN RTX 24GB). Network parameters were iteratively updated using the Adam optimizer with the default parameters. During the model training, the learning rate was set to 0.001 at the beginning, and 10 epochs later, it was reset to 0.0001, similarly, the number of selected local frame was initialized as the length of the input sequence  $T$  and then set to fixed value  $D$  after initial training (about 5 epochs). For the input CEUS video data, we first adopted an efficient temporal redundancy elimination strategy (Detailed as follows) and sampled  $T = 20$  frames uniformly as the input contrast sequence. The batch size is set to 10 and the number of maximal training epochs is set to 50.

To facilitate the model training under the limited CEUS data, there are several prerequisite steps required, which are detailed as follows. On the one hand, we perform the spatial-temporal pruning via lesion detection and temporal redundancy elimination; On the other hand, on-the-fly data augmentation and model pre-training are used to increase data diversity and speed up convergence.

**1) Motion Correction:** One of the primary challenges to monitor enhancement patterns through CEUS cine-loops is irregular lesion motion, which might be caused by numerous unpredictable factors, such as free breathing, cardiac movements, and emotional fluctuation, etc. [48]. Due to dramatic intensity variations over the examination, we choose the point-based registration techniques (PBRTs) with the feature descriptor, compact and real-time descriptors (CARD) [49], to align successive contrast frames.

**2) Temporal Redundancy Elimination:** In contrast to high temporal resolution of CEUS videos, dynamic enhancement patterns appear to vary slowly. Therefore, we adopt an inter-frame difference-based method [50] to eliminate temporal redundancy. Specifically, informative frames are defined as those points with significant pattern variation within a local window. Thus, we compute the inter-frame histograms difference  $\Delta f$  within a small interval  $2w + 1$  ( $w$  is set to 2) for each frame, and sort them in the descending order. Then, we preserve the former 48 frames to represent the original CEUS video.

**3) Lesion Detection:** In addition to redundancy elimination in time, lesion detection can be considered as an efficient way to reduce redundancy in space. As described in Section III-A.1, CEUS-Net from our previous work [43] is fine-tuned for nodule region localization. After motion correction and temporal pruning, 20 contrast frames are uniformly sampled from the preserved contrast subsequence as the network input. The spatial resolution of cropped contrast-specific view is



$360 \times 400$  pixels. In current implementation, we randomly sampled 40% of training set used in HiTAN evaluation as input. The effectiveness of our CEUS-Net on lesion detection and impact on final lesion recognition is evaluated by comparing with other detection models, with experimental results and discussions presented in Fig. S2 and Section C in Supplementary Materials.

**4) Data Augmentation:** In order to reduce the over-fitting risk, we augment the diversity of training samples with two main strategies, 1) randomly rotating or flipping the whole input contrast sequence; 2) randomly shifting the initial temporal points at the range  $[0, 10]$  for the uniform sampling of  $T = 20$  frames from the preserved 48 frames. Notably, this strategy was specifically designed for dynamic CEUS sequences.

**5) Model Pre-Training:** Model Pre-training is another effective strategy to avoid the risk of overfitting. To this end, we adopt the convolution auto-encoder<sup>2</sup> [51] to learn initial weights of enhancement representation module, where the original backbone is treated as the encoder and a symmetric decoder is appended to reconstruct CEUS frames. During the course of model training, the learning rate of the former convolutional layers is set to a relatively smaller value, about 1/10 of other parameters.

#### IV. EXPERIMENTS AND ANALYSES

In this section, we first describe the evaluated thyroid nodule dataset. Then, we compare the proposed HiTAN with several state-of-the-art methods, and validate the effectiveness of the important components of our method, including the hierarchical recognition mechanism and the local-to-global temporal fusion operator. After that, we further analyze the influence of several hyper-parameters as well as classification performances when fewer nodule types are considered. Finally, we verify the discriminative frames automatically identified by our HiTAN method. The classification accuracy of our model was evaluated by the class-specific sensitivity (SEN), mean accuracy (ACC), macro Precision, macro Recall and macro F1-score.

##### A. Dataset

Our Thyroid nodule dataset is a collection of 325 consecutive patients (Gender, F/M= 75/250; Age,  $45.8 \pm 12.3$  years) attending Nanjing Drum Tower Hospital from September 2016 to July 2018 for thyroid nodules examination. There are total 336 lesions with four pathological types incorporated in our dataset, including two types of benign nodules, i.e., 77 Nodular Goiter, 84 Adenoma and the malignant nodules, i.e., 101 Papillary thyroid carcinoma (PTC), 74 Papillary thyroid microcarcinoma (PTMC), the average maximum diameter of nodules is  $16.07 \pm 7.73$ mm. From the perspective of nodule pathological definition, Papillary microcarcinoma refers to the carcinoma smaller than 10mm, and their enhancement patterns generally show much less diversity than common Papillary carcinoma. All patients were examined by an expert

radiologist with over 10-year clinical experience. Examinations were performed on a Logiq E9 ultrasound scanner (GE Healthcare, Milwaukee, WI, USA) with the second-generation microbubble contrast SonoVue at low mechanical index (set below 0.12). Pathologies diagnosis of all cases were confirmed by biopsy or surgical specimens. Each video lasts around 3 minutes with a framerate of 15 fps and the spatial resolution of dual-view imaging is  $600 \times 800$  pixels. Approval was obtained by the ethics review board of local hospital and the informed consent was obtained from patients before this study.

##### B. Competing Methods

The proposed HiTAN method was first compared with the conventional parameters-based methods, where SVM classifiers are trained with kinetic parameters extracted from regional TICs [10]. Then, HiTAN was further compared with multi-view learning methods using handcrafted features, including multiple kernel learning (MKL) [52] and deep canonical correlation analysis (DCCA) [16]. Besides, we also compare our model with several state-of-the-art deep models in field of video recognition.

**1) Parameters-Based Methods:** As described in [10], there are two types of ROIs (i.e., tumor ROI  $R_t$  and parenchyma ROI  $R_p$ ) outlined in standard frame. Given the detected ROI  $R$ ,  $R_t$  and  $R_p$  are obtained by shrinking or enlarging  $R$  by a small factor respectively, and then  $R$  is subtracted from the  $R_p$ . In this way, not only the TIC parameters of lesion interior and surrounding normal tissues but also their relative differences are used to describe the perfusion process (28 kinetic features included in total). Furthermore, kinetic features are partitioned by temporal phases to train SVM classifiers so that the contributions of distinct phases can be exploited separately. SVM-A and SVM-A-P are independently trained using functional parameters from arterial and arterial-portal phases, respectively. In this study, LibSVM [55] is used as the library. We use the linear kernel and the penalty terms  $C$  of SVM-A and SVM-A-P is selected by grid search from  $\{0.1, 0.2 \dots, 10\}$  via cross-validation.

**2) Multi-View Fusion Methods:** In line with [16], temporal dynamics are represented using three representative contrast frames of different temporal phases. 66-dimensional statistical texture features (e.g., Gray-level co-occurrence matrix, GLCM and Local phase, LP) are first extracted using Pyradiomics package<sup>3</sup> to represent the appearance of enhancement for each frame. After that, three view features are fused by multi-view learning algorithms (MKL [52] and DCCA [16]) for lesion recognition. As for DCCA, the generated six-view features are fed into the multi-kernel SVM classifier to boost the diagnostic accuracy. The implementation details of MKL and DCCA are presented in Section A of Supplementary Material.

**3) Deep Learning-Based Methods:** Different from conventional methods, deep models take sequential contrast frames as inputs and unify dynamic feature extraction and classifier construction into an end-to-end framework. The key challenge of CEUS video classification is how to model the temporal

<sup>2</sup><https://github.com/ShayanPersonal/stacked-autoencoder-pytorch>

<sup>3</sup><https://pyradiomics.readthedocs.io/en/latest/>

TABLE I

CLASSIFICATION PERFORMANCES OF DIFFERENT COMPETING METHODS FOR THE THYROID NODULE DIAGNOSIS. WE REPORTED THE SENSITIVITY OF EACH NODULE TYPE AND FOUR MULTI-CLASS METRICS, INCLUDING THE MEAN ACCURACY, MACRO PRECISION, MACRO RECALL AND MACRO-F1 SCORE (%). THE TERMS A AND B IN "A $\pm$ B" DENOTE THE MEAN AND STANDARD DEVIATION

Method	Thyroid Nodules							
	SEN. Nodular Goiter	SEN. Adenoma	SEN. PTC	SEN. PTMC	ACC	Precision	Recall	F1-score
SVM-A [10]	40.27 $\pm$ 6.10	41.41 $\pm$ 4.55	69.90 $\pm$ 5.17	80.00 $\pm$ 5.33	58.47 $\pm$ 2.83	59.06 $\pm$ 3.23	57.90 $\pm$ 2.84	57.25 $\pm$ 2.94
SVM-A-P [10]	47.47 $\pm$ 5.44	47.06 $\pm$ 4.71	59.43 $\pm$ 5.57	86.40 $\pm$ 4.80	59.65 $\pm$ 2.76	61.74 $\pm$ 2.96	60.09 $\pm$ 2.67	59.29 $\pm$ 2.80
MKL [52]	54.13 $\pm$ 5.44	52.71 $\pm$ 4.85	69.71 $\pm$ 5.37	72.80 $\pm$ 5.94	62.71 $\pm$ 2.75	62.80 $\pm$ 2.93	62.34 $\pm$ 2.77	62.20 $\pm$ 2.79
DCCA [16]	61.07 $\pm$ 5.87	58.82 $\pm$ 4.99	56.76 $\pm$ 3.79	89.93 $\pm$ 3.20	65.76 $\pm$ 2.30	68.40 $\pm$ 2.16	66.90 $\pm$ 2.29	66.13 $\pm$ 2.35
C3D [27]	40.53 $\pm$ 5.31	54.35 $\pm$ 4.79	80.95 $\pm$ 4.67	75.73 $\pm$ 3.20	64.24 $\pm$ 1.75	64.07 $\pm$ 1.86	62.89 $\pm$ 1.74	62.67 $\pm$ 1.85
R2Plus1D [29]	58.40 $\pm$ 5.09	70.12 $\pm$ 4.38	58.86 $\pm$ 5.37	<b>90.93<math>\pm</math>3.20</b>	68.65 $\pm$ 2.59	71.15 $\pm$ 2.48	69.58 $\pm$ 2.52	68.72 $\pm$ 2.61
TCN [53]	45.33 $\pm$ 5.31	<b>76.47<math>\pm</math>4.55</b>	61.52 $\pm$ 4.24	90.11 $\pm$ 3.33	67.94 $\pm$ 1.86	70.62 $\pm$ 1.96	68.31 $\pm$ 1.86	67.02 $\pm$ 2.03
CNN-LSTM [31]	58.67 $\pm$ 5.66	64.47 $\pm$ 4.85	72.01 $\pm$ 3.64	90.90 $\pm$ 3.27	71.35 $\pm$ 2.21	73.03 $\pm$ 2.20	71.52 $\pm$ 2.28	71.18 $\pm$ 2.32
TSN [54]	52.27 $\pm$ 5.86	69.88 $\pm$ 5.08	77.14 $\pm$ 4.04	89.60 $\pm$ 3.31	72.59 $\pm$ 1.81	73.97 $\pm$ 2.07	72.22 $\pm$ 1.93	71.86 $\pm$ 1.91
TSM [39]	58.93 $\pm$ 5.23	71.76 $\pm$ 4.71	80.38 $\pm$ 4.11	89.07 $\pm$ 3.20	75.41 $\pm$ 2.41	76.36 $\pm$ 2.44	75.04 $\pm$ 2.41	74.91 $\pm$ 2.46
HiTAN	<b>68.27<math>\pm</math>4.93</b>	76.24 $\pm$ 4.40	<b>84.38<math>\pm</math>3.93</b>	90.40 $\pm$ 3.30	<b>80.18<math>\pm</math>2.36</b>	<b>80.83<math>\pm</math>2.42</b>	<b>79.88<math>\pm</math>2.42</b>	<b>79.90<math>\pm</math>2.44</b>

structure of perfusion sequence. According to the way of dynamics learning, we divided these deep models into two categories: 1) jointly modeling spatial-temporal variations within adjacent frames using 3D convolution, such as C3D [27] and R2Plus1D (Pseudo 3D) [29]; 2) learning sequence-level representation by aggregating frame-level temporal patterns, including TCN [53], CNN-LSTM [31], TSN [54] and TSM that shifts temporal dimensions for information exchange within neighboring frames [39]. Similarly, the training steps of all video classification methods are detailed in Table S1 of the Supplementary Material.

In our experiments, we adopt a 5-fold cross-validation strategy for performance evaluation of all methods. First, our available CEUS videos are split into five equal-size folds by preserving the percentage of samples for each class. Then, one fold is held out as the testing set, while the remaining folds are randomly split into training and validation sets in the ratio of 4: 1. By repeating this process five times, we could obtain the mean score and standard deviation over five folds. Notably, the implementation of these competing methods shares the same lesion detection results with our HiTAN, reducing the impact of perfusion region selection on classification performance. In order to avoid model overfitting, both for 3D and 2D convolutional filters, initial weights are learned by 3D(2D) convolutional auto-encoders. As for the binary-classifier (i.e., SVM, MKL), we adopt the one-vs-rest binary classification approach to perform the multi-class classification.

### C. Analysis of Diagnostic Performance

In this group of experiments, different types of classification models are used for performance comparison. Table I presents the quantitative classification results obtained by the competing methods as well as our proposed HiTAN.

Some important observations can be summarized from this table. First, deep learning-based methods (i.e. I3D, R2Plus1D, TCN and HiTAN) generally yield better recognition results than traditional methods either based on handcrafted

features (i.e. MKL and DCCA) or TIC kinetic parameters (i.e. SVM-A and SVM-A-P). The highest mean accuracy is observed for our HiTAN with 80.18%, even for the baseline 3D CNN (C3D) that also achieves at least performance with 64.24%, slightly lower than DCCA with 65.76%. This indicates that, effective enhancement patterns representation (both for a single frame or local temporal window) lay the foundation for subsequent temporal fusion and prediction. And the task-oriented convolutional features are superior to pre-defined statistical texture features in terms of diverse enhancement appearances. In contrast, limited kinetic parameters derived from TICs is the weakest, especially for the differentiation of benign nodules. This shows that simply averaging pixel intensities inevitably loses the information of contrast distribution, such as the interior heterogeneity or peripheral ring-enhancement. Second, as for C3D, the performance improvement is less than R2Plus1D and other temporal fusion models (with mean accuracy  $\geq 65\%$ ), which could be explained by the parameter overheads of 3D convolutional filters even with model pre-train. Due to the relatively small number of CEUS data compared with high-dimensional inputs, leveraging 2D CNN to extract image-level features and then fusing temporal information is more appropriate since the largely reduced parameters size.

Third, we could observe that recognizing different types of nodules has different requirements for the local or global temporal dynamics fusion. For a better understanding, we decompose the temporal dynamics into two levels, that is, short-range features (e.g., the degree of enhancements or homogeneity presented in several disjoint frames) and long-range features (e.g., the centripetal or eccentricity enhancements, fast (slow) wash-in or out only observed in local temporal windows). Taking the papillary thyroid microcarcinoma (PTMC) as example, most models with different temporal modeling mechanisms can achieve the sensitivity over 85%, since the typical homogeneous low-enhancement is easy to be captured at both two levels. And for benign nodules, the mean sensitivity of Adenoma is higher than that of Nodular



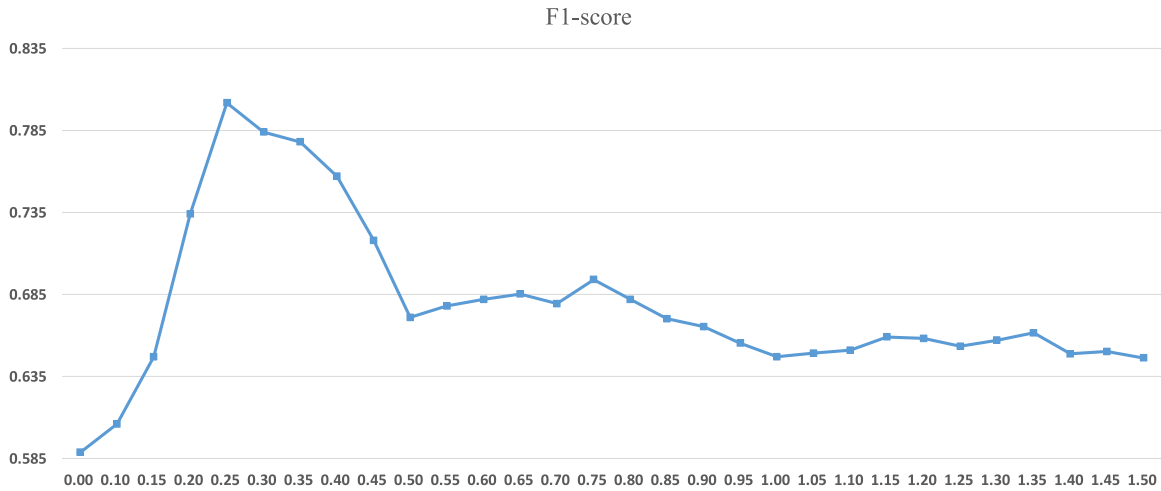


Fig. 4. F1-score curves obtained by our HiTAN method by changing the trade-off parameter  $\lambda_d$  from 0. to 1.5.

Goiter, which can also be explained by its characteristic rim hyper-enhancement and relatively slow fade out. Compared with the competing methods, our proposed HiTAN model significantly outperforms other models in the more challenging classification task of Nodular Goiter and papillary thyroid carcinoma (PTC) that show more diverse and complicated enhancement patterns, with the sensitivity of 84.38% and 68.27%. Experimental results validate that, as a local-to-global temporal modeling manner, LGTA is more suitable to capture the effective dynamic features by embedding the temporal contextual information into each identified key point. In comparison, the sensitivity of PTC observed for TCN and CNN-LSTM are only 61.52% and 72.01% and the results of Nodular Goiter for TCN and TSN are below 55%. It is perhaps due to that these methods lack the attention mechanism that selectively fuse partial representative patterns, especially for CNN-LSTM, the internal gating mechanism is hard to preserve the valuable characteristics if they appear in the early perfusion stage. Meanwhile, we found that TSM only behaves inferior to our HiTAN, possibly due to the sufficient temporal interaction by exchanging features within neighboring frames with the minimal parameters cost.

#### D. Effectiveness of Hierarchical Recognition Mechanism

As introduced in Section III-A.2, a key component of our proposed method was the hierarchical lesion recognition module which decomposed the diagnosis of nodules into two stages, thereby taking the diagnostic hierarchy into consideration. To evaluate the effectiveness of this hierarchical recognition mechanism, we design another version of our method (i.e., Non-local temporal aggregation network (NL-TAN)) for comparison, in which pathological prediction is made at one step by only preserving the NLTA operator for enhancement information fusion (Detailed network structure is illustrated in Fig. S3 of the Supplementary Material). Besides, we fine-tune the trade-off parameter  $\lambda_d$  from 0 to 1.5, aiming at evaluating the relative importance of classification tasks at two stages.

In our experiments, the mean accuracy and F1-score achieved by NL-TAN is 0.7388 and 0.731, respectively. The class-specific sensitivity scores of Nodular Goiter and Adenoma drop to 0.595 and 0.689 compared with our HiTAN method with 0.683 and 0.762, which demonstrates the effectiveness of the GRUs-based knowledge propagation between the two-step recognition tasks. Nonetheless, compared with results shown in Table I, overall classification performances are higher than other models, which further verify the superiority of non-local fusion manner compared with the LSTM and temporal convolution(pooling). We consider that establishing one-to-one relation between any two temporal points can be understood as an imitation of the radiologists' back-and-forth observation procedure to some extent, which is more beneficial to extract diverse perfusion patterns.

As illustrated in Fig. 4, the classification performance in terms of F1-score shows a significant upward trend when  $\lambda_d$  rises from 0 to 0.25 and peaked at 0.802. Then, it goes through a considerable and rapid decline with a proportion of 16.3% until  $\lambda_d$  reaches to 0.5. After that, F1-score continues to decrease at a slower rate with few fluctuations, and is steadily to approximately 0.65. The overall changing trend clearly shows that characterization-level classification task always plays a dominant role in our hierarchical nodule recognition network. That is, distinguishing specific pathological type deserves a higher weight to capture subtle enhancement characteristics with strong discriminative capacities. Meanwhile, we observe that there was a dramatic performance drop when  $\lambda_d$  is too small (below 0.15). Especially when  $\lambda_d$  reduces to zero, the lowest value of F1-score is observed with 0.588. A direct reason is that the lack of coarse-grained label information would cause the differentiation-level representation  $h_1$  and  $o_1$  no longer encode enhancement dynamics with desired discriminant power, leading a negative impact on key frame identification and local-to-global temporal fusion in LGTA. Another interesting phenomenon is that NL-TAN still yields superior diagnostic performance compared with the case of  $\lambda_d = 0$  in our experiment. This can be attributed to the task-oriented non-local temporal fusion and fewer network

parameters by removing RNN-based hierarchical classification mechanism under the limited CEUS samples. When  $\lambda_d$  exceeds 0.7, we found that performance degradation become relatively steadily compared with  $\lambda_d \in [0.35, 0.5]$ . It is perhaps due to that certain nodule types (i.e., Adenoma and PTMC) are easy to distinguish, which have less strict requirement for local-to-global enhancement feature fusion. In conclusion, the optimal weight value of  $\lambda_d$  is dependent on to-be-classified nodule types as well as the sample size. In current implementation, we empirically fixed the trade-off parameter  $\lambda_d$  to 0.25.

### E. Effectiveness of Local-to-Global Temporal Fusion

As introduced in III-A.2, the central idea of LGTA is to identify part of discriminative frames encoding salient enhancement patterns first, and then embed global enhancement dynamics into each of them. In this way, *Local-to-Global* enhancement features can be comprehensively integrated to characterize intra-nodular vascularity.

To evaluate the effectiveness of this fusion mechanism, we also design three other variants, 1) *Local-to-Local Temporal Aggregation (LLTA)*, where inter-frame interactions are only inferred within local salient enhancement patterns; 2) *Global-to-Local Temporal Aggregation (GLTA)*, enhancement characteristics from identified keyframes  $\mathbf{F}$  are incorporated into each enhancement descriptor  $v_t$ , rather than embedding global features  $\mathbf{G}$  into identified key patterns  $\mathbf{F}$ ; 3) *Global-to-Global Temporal Aggregation (GGTA)*, where the step of temporal weighting estimation is removed, reducing to a non-local temporal aggregation (NLTA) of the whole enhancement descriptors  $\mathbf{G}$ . Since that hierarchical classification mechanism is preserved in these variants, the corresponding hierarchical convolution networks are named as LL-HCN, GL-HCN, and GG-HCN, respectively. Detailed structures are illustrated in Fig. S4 and S5 in the Supplementary Materials.

The corresponding experimental results are presented in Table II, from which we could have at least two observations. 1) Compared with the other two variants that only focus on either local or global temporal information (i.e., LL-HCN and GG-HCN), GL-HCN and LG-HCN (our HiTAN method) achieve better or competitive classification performances. This implies that, the manner of local and global temporal modeling is more effective to fuse discriminative enhancement characteristics regarding vessel distribution or blood hemodynamics. As in the case of Nodular Goiter classification, the manner of local-to-local temporal fusion shows a significant performance drop in comparison to other three manners that at least have a global temporal receptive field, demonstrating that more discriminative characteristics of Nodular Goiter might consist in the overall enhancement tendency; Also, the sensitivity scores of PTC observed for GG-HCN are slower, which might be explained by that typical heterogeneous enhancement patterns often appear around the perfusion peak while the wash-out phase tends to be relatively homogeneous, thus a simple global fusion might be suboptimal; 2) Our HiTAN method outperforms GL-HCN on three nodule types except for Adenoma, for example, the mean sensitivity for Nodular

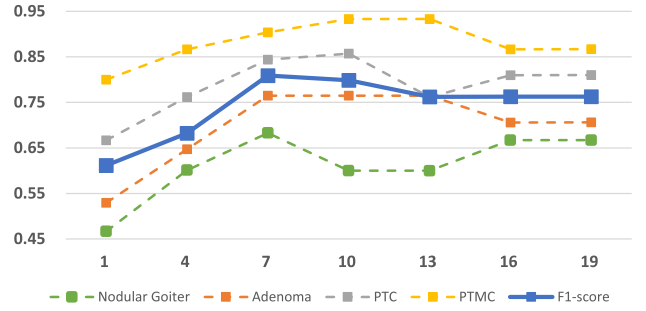


Fig. 5. Results of thyroid nodule classification obtained by our HiTAN method by tuning the numbers of selected local key frames (i.e.,  $D$  from  $\{1, 4, 7, 10, 13, 16, 19\}$ ). Dotted lines plot the class-specific sensitivity scores and the solid line corresponds to the F1-score.

Goiter is 0.683 vs. 0.664 and for PTMC is 0.843 vs. 0.817, while for Adenoma classification is 0.762 vs. 0.774. The slight performance improvement achieved by GL-HCN might be attributed to the relatively typical enhancement patterns of Adenoma, e.g., ring-like hyper-enhancement that appears across the majority of perfusion process.

### F. Influence of the Number of Selected Key Frames

As introduced in LGTA module, partial key frames presenting typical enhancement patterns are explicitly extracted with the help of temporal attention score  $\mathbf{A}$ . In this group of experiments, we investigated the influence of the number of selected frames (denoted as  $D$ ) on the final classification performances. Specifically, we orderly selected  $D$  from  $\{1, 4, 7, 10, 13, 16, 19\}$  in HiTAN and presented the corresponding results (evaluated by class-specific sensitivity and F1-score) in Fig. 5.

As shown in Fig. 5, both class-specific sensitivity and F1-score increase rapidly when choosing  $D$  from 1 to 7 for input CEUS sequences with the length  $T = 20$ . Especially for Nodular Goiter, its sensitivity score is even lower than 0.5 when setting  $D$  to 1. In contrast, the impact of the number of key frame on the recognition of PTMC is much smaller. This implies that two few contrast frames are not enough to encompass sufficient contrast information when identifying some types of nodules with diverse and complicated dynamic variations. Besides, we also observe that the performance is relatively steady when  $D$  is changed from 10 to 19, the overall F1-score decreased by a small range from 0.788 to 0.763 since the influence of  $D$  on the diagnosis performance varies with the nodule type. As can be seen, only small number of frames are required to recognize the PTMC accurately for its apparent focal low-enhancement feature, while for Adenoma and PTC,  $[7, 10]$  is optimal interval to capture their representative enhancement patterns. As for Nodular Goiter, it is more difficult to deduce the varying trend roughly from the green curve since its more complex perfusion characteristics. If a larger thyroid dataset could be collected in the future, the corresponding impact might be further explored. In our implementation,  $D$  is chosen as 7 to largely incorporate potentially informative local temporal points, as well as to

TABLE II

CLASSIFICATION PERFORMANCES OF DIFFERENT VARIANTS OF LGTA OPERATOR. WE REPORTED THE SENSITIVITY OF EACH TYPE, MEAN ACCURACY, AND MACRO F1-SCORE (%). THE TERMS A AND B IN “A $\pm$ B” DENOTE THE MEAN AND STANDARD DEVIATION

Method	SEN. Nodular Goiter	SEN. Adenoma	SEN. PTC	SEN. PTMC	ACC	F1-score
LL-HCN	57.87 $\pm$ 5.24	74.35 $\pm$ 4.67	76.05 $\pm$ 3.92	87.47 $\pm$ 5.44	74.12 $\pm$ 2.46	73.62 $\pm$ 2.57
GG-HCN	67.47 $\pm$ 5.31	71.06 $\pm$ 4.68	79.24 $\pm$ 3.98	82.93 $\pm$ 3.31	75.82 $\pm$ 2.52	74.96 $\pm$ 2.63
GL-HCN	66.42 $\pm$ 5.16	<b>77.41<math>\pm</math>4.90</b>	81.71 $\pm$ 3.97	87.20 $\pm$ 4.96	78.47 $\pm$ 2.35	78.11 $\pm$ 2.32
LG-HCN (HiTAN)	<b>68.27<math>\pm</math>4.93</b>	76.24 $\pm$ 4.40	<b>84.38<math>\pm</math>3.93</b>	<b>90.40<math>\pm</math>3.30</b>	<b>80.18<math>\pm</math>2.36</b>	<b>79.90<math>\pm</math>2.44</b>

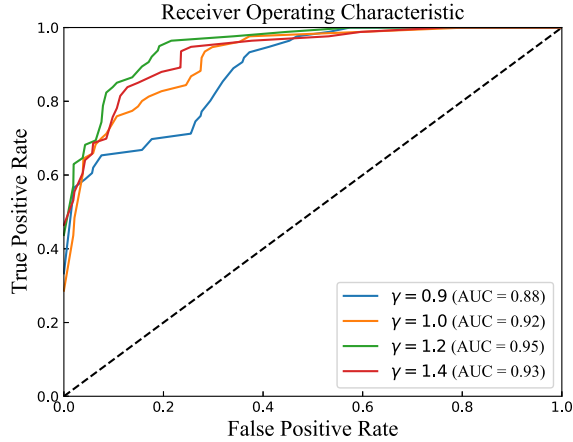


Fig. 6. Multi-class ROC curves obtained by our HiTAN method by changing the enlarging factor  $\gamma$  from {0.9, 1.0, 1.2, 1.4}. Curves of different colors correspond to different degrees of surrounding tissues incorporated for the pattern representation.

reduce the computational complexity and memory cost during the training.

#### G. Influence of the Size of Incorporated Tissues

As known from qualitative perfusion descriptions, such as hyper-enhancement, iso-enhancement and hypo-enhancement, contrast concentration of surrounding tissues provides necessary reference to measure the blood supply within lesions. In previous implementations, we enlarged the bounding-box by the fixed factor  $\gamma = 1.2$ . To evaluate the influence of the size of included surrounding tissues. In this group of experiments, we train the network with the  $\gamma = 0.9, 1.0, 1.2, 1.4$ , respectively. And we evaluate the impact on the classification performance by the multi-class ROC curve calculated by averaging the class-specific ROC curves of four pathological types.

From Fig. 6, we can see that the size of included surrounding tissues is closely correlated with the classification performance of our HiTAN method, as different curves (corresponding to different  $\gamma$  value) separate from each other clearly. For different  $\gamma$ , we can make conclusion as follows, 1) only the contrast appearance of lesion interior and boundary is not enough to depict the complicate patterns with  $AUC = 0.88$  and  $AUC = 0.92$  when  $\gamma$  is set to 0.9 or 1.0. It may be because interior contrast concentration variation could only reflect the local blood supply information (i.e., the degree of homogeneity) while the feature of degree of enhancement (e.g., iso-enhancement, rim-enhancement) is inevitably discarded. 2) the performance of using larger  $\gamma = 1.4$  also

slightly decreased, this result is as expected because excessive incorporated surrounding tissues also include potentially uninformative or even disturbing perfusion features for nodule classification. Therefore, we chose an intermediate value 1.2 in our current implementation.

#### H. Different Combinations of Four Nodule Types

In order to further evaluate the efficacy of our HiTAN method in pathological enhancement characteristics mining, we added another six groups of experiments in which fewer nodule types are considered. The combinations of different nodule types and their respective classification results are reported in Table III.

Several observations can be summarized from Table III.

1) Compared with the recognition accuracy of other three types, the class-specific sensitivity of Nodular Goiter is still the lowest in different groups of experiments, even in the case of binary classification. Similar results can also be found in Table I and II, which might be attributed to larger enhancement pattern variations correlated with complex vascular structures. Meanwhile, the increased perfusion diversity together with relatively small number of training samples also aggravate the difficulty of capturing pathological-specific enhancement characteristics. While for the other benign type Adenoma, it seems to be easier with the lowest sensitivity of 0.788 that outperforms the best result of Nodular Goiter; 2) The highest classification performance is achieved by the group of Adenoma vs. PTMC in terms of class-specific sensitivity observed with 0.835 for Adenoma and 0.946 for PTMC, and mean accuracy observed with 0.887, considering their distinctly distinguished perfusion patterns. As observed from identified key patterns, ring-like hyper-enhancement with a clear margin provides a strong clue for Adenoma while homogenous hypo-enhancement offers an evident indication for PTMC; 3) We find that fine-tuning the trade-off parameter  $\lambda_d$  with a gap of 0.05 is necessary for different tasks. Especially for those binary classification tasks where contrast appearances of different types might overlap to some degree, a larger  $\lambda_d$  compared with 0.25 set in the four-class classification task is more reasonable, since the two-stage tasks share the same label information and a larger weight could ensure that a discriminative global representation can be learned at the coarse-grained stage in these tasks.

#### I. Automatically-Identified Discriminative Subsequence

In our HiTAN method, dynamic enhancement characteristics were fused by our proposed local-to-global temporal



TABLE III

CLASSIFICATION RESULTS OF SIX GROUPS OF EXPERIMENTS OF DIFFERENT COMBINATIONS OF NODULE TYPES. WE REPORTED THE SENSITIVITY OF EACH TYPE, MEAN ACCURACY, MACRO PRECISION, MACRO RECALL, AND MACRO F1-SCORE (%). THE TERMS A AND B IN " $A \pm B$ " DENOTE THE MEAN AND STANDARD DEVIATION

No.	Nodule types	$\lambda_d$	Sen. Nodular Goiter	Sen. Adenoma	Sen. PTC	Sen. PTMC	ACC	Precision	Recall	F1-score
1	Nodular Goiter vs. PTC	0.65	62.67 $\pm$ 9.98	-	68.57 $\pm$ 2.33	-	69.44 $\pm$ 5.27	68.50 $\pm$ 5.61	68.48 $\pm$ 5.93	68.41 $\pm$ 5.81
2	Nodular Goiter vs. PTMC	0.60	69.33 $\pm$ 5.33	-	-	89.33 $\pm$ 3.27	79.33 $\pm$ 3.88	80.59 $\pm$ 3.80	79.33 $\pm$ 3.89	79.11 $\pm$ 3.95
3	Adenoma vs. PTC	0.50	-	85.88 $\pm$ 2.88	77.14 $\pm$ 2.10	-	81.05 $\pm$ 1.97	81.20 $\pm$ 2.02	81.51 $\pm$ 2.03	81.02 $\pm$ 1.98
4	Adenoma vs. PTMC	0.35	-	83.53 $\pm$ 2.35	-	94.67 $\pm$ 2.67	88.75 $\pm$ 2.50	89.10 $\pm$ 2.51	88.94 $\pm$ 2.53	88.75 $\pm$ 2.59
5	Nodular Goiter vs. PTC vs. PTMC	0.35	64.67 $\pm$ 7.22	-	77.14 $\pm$ 4.10	89.39 $\pm$ 2.99	77.67 $\pm$ 4.76	78.84 $\pm$ 4.74	77.75 $\pm$ 4.46	77.57 $\pm$ 4.69
6	Adenoma vs. PTC vs. PTMC	0.30	-	78.82 $\pm$ 2.88	78.10 $\pm$ 5.71	88.16 $\pm$ 2.67	81.13 $\pm$ 3.58	81.51 $\pm$ 3.42	81.64 $\pm$ 3.37	81.24 $\pm$ 3.52

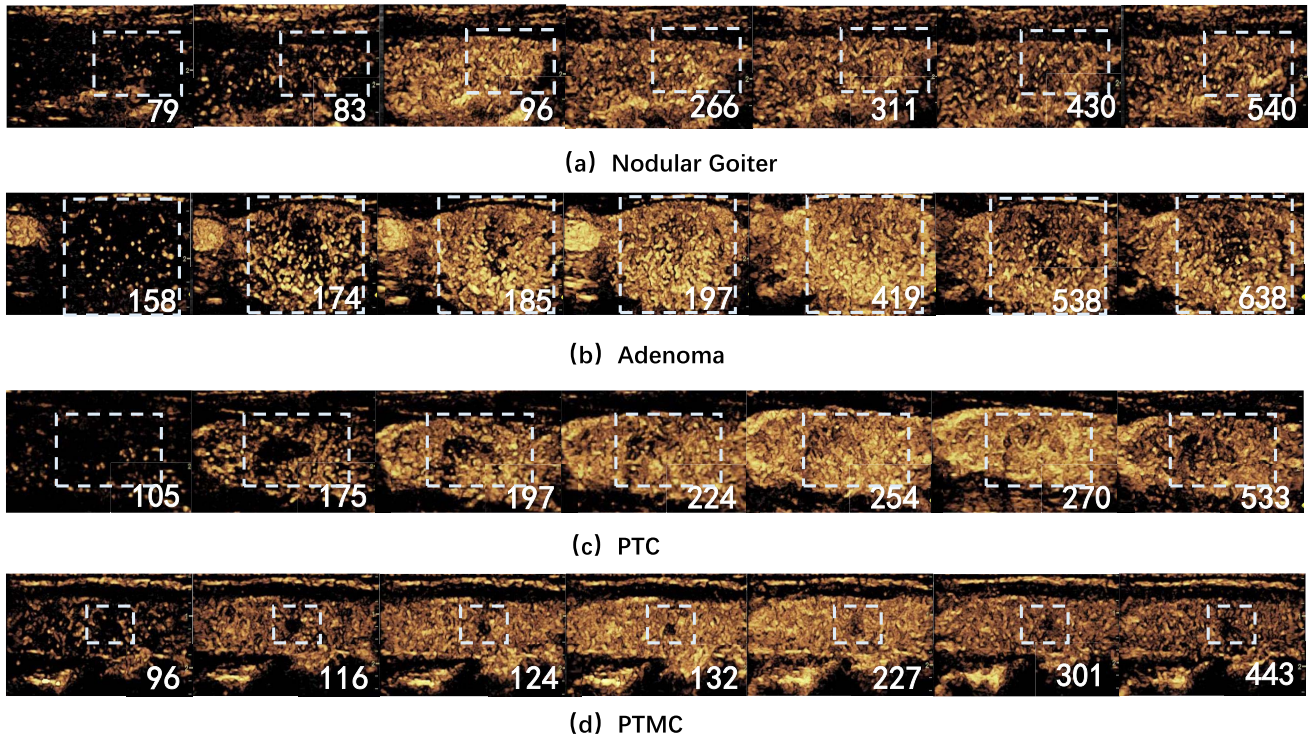


Fig. 7. Discriminative temporal locations automatically identified by our proposed method. Each row corresponds to one specific nodule type. For a better visual illustration, we cropped the original contrast view and outlined the delineated contrast regions using white dotted rectangles.

aggregation module, which can automatically identify those representative contrast frames as local key points for the class-specific perfusion feature extraction. In Fig. 7, we visually verify those automatically-identified temporal points in distinguishing thyroid nodules.

Specifically, Fig. 7 presents the temporal supports identified by the HiTAN method for each thyroid nodule type (i.e., each row corresponds to one specific type). For a better illustration of salient enhancement patterns, we cropped the original contrast view and outlined the delineated contrast regions using white dotted rectangles. Besides, the corresponding temporal index is also appended to each contrast frame. From Fig. 7, we observed that the perfusion patterns contained in identified temporal points across distinctive phases are consistent with previous clinical studies [9], [56], which confirms the ability of our method in localizing key frames. As for PTMC (shown in the last row), the typical focal low-enhancement can be

clearly observed from each selected frames, which implies the less requirements in accurate key points localization and interprets the basically same sensitivity results of different ablation experiments. While for PTC (shown in the third row), concentrating on representative frames is relatively more important, since the typical concentric enhancement pattern is distributed over distant temporal points. We could hypothesize that if more attention is paid on neighboring points 254 to 270, this network could be misguided by the seemingly homogenous enhancement characteristic. In addition, attention mechanism also enables to preserve the fading frame even in case of a gradually wash-out process. Similar observations can also be found for benign nodules, where the homogenous hyper-enhancement as well as fast wash-in (out) characteristic is preserved for the Goiter Nodular, while the rim hyper-enhancement as well the fast wash-in, slow wash-out is preserved for Adenoma, respectively.

TABLE IV

A BRIEF DESCRIPTION OF THE STATE-OF-THE-ART STUDIES USING DYNAMIC CONTRAST-ENHANCED ULTRASOUND IMAGING.  
SEN DENOTES THE SENSITIVITY AND SPE DENOTES THE SPECIFICITY

Reference	Imaging	Contrast agent	Number of nodules (Benign/Malignant)	Feature	Method	class	Performance
Zhou X <i>et al.</i> [57]	CEUS	2.4mL SonoVue	68 /93	Six functional features from TIC	Univariate logistic regression	2	SEN: 80.41% and SPE: 80%
Zhao H <i>et al.</i> [58]	CEUS	2.4mL SonoVue	57 /60	Eight qualitative enhancement patterns	Logistic regression analysis	2	SEN: 89.47% and SPE: 88.33%
He Y <i>et al.</i> [59]	CEUS	1.8mL SonoVue	59 / 29	Six functional features from TIC	Student's t-test	2	SEN: 79.3% and SPE: 91.5%
Acharya <i>et al.</i> [60]	3D CEUS	2.5mL SonoVue	10 /10	Three texture features and Discrete Wavelet Transform features	K-nearest neighbors	2	SEN: 98% and SPE: 99.8%
Xi X <i>et al.</i> [61]	CEUS + Elastography	1.2 mL SonoVue	134 / 29	Three qualitative enhancement feature and firmness features	Student's t-test	2	SEN: 51.7% and SPE: 88.1%
Luo W <i>et al.</i> [62]	US + CEUS + Elastography + color Doppler US	1.2 mL SonoVue	101 / 220	Eighteen shape, functional, and vascular distribution features	Decision tree	2	SEN: 98.6% and SPE: 80.1%
Our method	CEUS	2.4mL SonoVue	161 /175	Deep convolution feature	HiTAN	4	SEN of four types: 68.27%, 76.24%, 84.38%, and 90.40%

## V. DISCUSSION

In this section, we first summarize the main differences between our proposed HiTAN method and previous studies on CEUS-related tumor diagnosis. We also point out the limitations of our proposed method as well as potential solutions to deal with these limitations in the future.

### A. Comparison With Previous Studies

Compared with the conventional TIC analysis [10], [25] and multi-view fusion methods [16], [17], our proposed HiTAN method accepts dynamic CEUS sequences as inputs to develop an end-to-end hierarchical classification model. Capitalizing on task-oriented, hierarchical feature representation learning, this method is able to quantify diverse perfusion patterns with enhanced description power. Apart from that, different from general CNN-based video classification models [27], [31], [39], [54], another major advantage is the improved model interpretability, which consists in the incorporated hierarchical nodule classification mechanism and the designed local-to-global temporal fusion module that models a comprehensive analysis of local enhancement patterns (e.g., ring-like enhancement) and global perfusion tendency (e.g., concentric or dispersed enhancement, fast wash-in or slow fade-out) in clinics. Particularly, the explicit identification of salient enhancement patterns helps to inform clinicians of the diagnosis basis of current prediction, which is beneficial to interpret the inner workings of our model and summarize typical perfusion patterns of diagnostic value.

In Table IV, we briefly summarize several notable work reported in the literature for thyroid nodules classification using dynamic CEUS imaging, including two statistical-based methods and four conventional learning-based methods (i.e., logistic regression analysis, K-nearest neighbors, and decision tree). It should be noted that a direct comparison with

these studies is impossible due to the utilization of datasets, including the concentration of contrast agent, the number of subjects, the number of classes, and the varying partition of training/testing set. However, we could still have several observations by rough comparison. *First*, our method was performed on a much larger cohort of 336 nodules covering four types, while the majority of methods in Table IV were evaluated on a relatively smaller dataset (less than 200 samples), which should be more convincing. *Second*, most methods still rest on univariable or multivariable analysis (e.g., qualitative enhancement pattern description or quantitative functional parameters from time-intensity curves). In spite of clear clinical definitions of these variables, their numerical recordings tend to be subjective. Few learning-based methods still depend on hand-crafted feature extraction from single CEUS image without data-driven enhancement pattern representations, let alone dynamic enhancement characteristics fusion. *Third*, existing methods mainly focus on the binary differentiation of benign and malignant nodules, while our method has exploited the more challenging task of pathological recognition, which incorporates a hierarchical lesion recognition mechanism.

### B. Limitations and Future Work

First, as pointed out by first-line radiologists, single-model CEUS imaging with vascular distribution information is still insufficient to comprehensively characterize nodules, and information from other US modalities (e.g., B-mode and elastography US with morphological and stiffness features) should be jointly considered. Also, a larger cohort of subjects could be included to ensure that the varying perfusion patterns of specific nodule type can be captured, thus further improving the prediction accuracy and reducing the overfitting risk. We expect that a large scale multi-modal thyroid US dataset (e.g., B-mode US, color Doppler US,

and elastography US, etc.) could be collected to overcome information limitation. By then, a multi-modal US analysis platform which combines this type of video classification methods and conventional radiomic analysis techniques might produce more reliable diagnosis results. Second, informative perfusion regions are delineated based on the segmentation mask output by CEUS-Net, and the bounding-box is enlarged by a fixed factor to incorporate the blood supply information of surrounding parenchyma. Though we have conducted a group of experiments to discuss the appropriate value of the enlarging proportion, it is still not flexible enough to set a fixed proportion for all CEUS sequences. If this parameter is set as learnable, i.e., the proportion of included normal tissues is adaptive to different subjects, dynamics learning capacity could be further facilitated. If our collected samples are sufficient enough, another feasible approach is to add a weakly-supervised detection branch to original classification network and employ the ROI-Pooling to focus on informative perfusion areas.

## VI. CONCLUSION

In this paper, a hierarchical temporal attention network (HiTAN) was proposed to model a coarse-to-fine diagnostic procedure, i.e., from coarse-grained (benign/malignant) differentiation to fine-grained (pathological type) characterization. In this way, a local-to-global temporal aggregation mechanism could be embedded to model the complicated temporal relation. On our collected thyroid nodule dataset, the effectiveness of our method on automatic enhancement patterns representation, temporal modeling and hierarchical dependency has been extensively evaluated. Compared with several state-of-the-art CAD methods, our proposed method has achieved better or at least comparable recognition performance. More importantly, our designed temporal fusion and hierarchical recognition mechanism take the dynamic enhancement characteristics full into account, providing a more intuitive understanding of model prediction.

## REFERENCES

- [1] P. Dijkmans *et al.*, "Microbubbles and ultrasound: From diagnosis to therapy," *Eur. J. Echocardiogr.*, vol. 5, no. 4, pp. 245–246, 2004.
- [2] K. Seitz *et al.*, "Contrast-enhanced ultrasound (CEUS) for the characterization of focal liver lesions in clinical practice," *Ultraschall Der Medizin*, vol. 31, no. 5, pp. 492–499, 2010.
- [3] I. Sporea, R. Sirli, A. Martie, A. Popescu, and M. Danila, "How useful is contrast enhanced ultrasonography for the characterization of focal liver lesions?" *J. Gastrointestinal Liver Diseases*, vol. 19, no. 4, pp. 1–6, 2010.
- [4] M.-X. Tang *et al.*, "Quantitative contrast-enhanced ultrasound imaging: A review of sources of variability," *Interface Focus*, vol. 1, no. 4, pp. 520–539, Aug. 2011.
- [5] S. Delorme and M. V. Knopp, "Non-invasive vascular imaging: Assessing tumour vascularity," *Eur. Radiol.*, vol. 8, no. 4, pp. 517–527, May 1998.
- [6] K. Wei, E. Le, J.-P. Bin, M. Coggins, J. Thorpe, and S. Kaul, "Quantification of renal blood flow with contrast-enhanced ultrasound," *J. Amer. College Cardiol.*, vol. 37, no. 4, pp. 1135–1140, Mar. 2001.
- [7] J. Folkman, "Role of angiogenesis in tumor growth and metastasis," *Seminars Oncol.*, vol. 29, no. 6, pp. 15–18, Dec. 2002.
- [8] G. Russo, M. Mischi, W. Scheepens, J. J. De la Rosette, and H. Wijkstra, "Angiogenesis in prostate cancer: Onset, progression and imaging," *BJU Int.*, vol. 110, no. 11c, pp. E794–E808, Dec. 2012.
- [9] J. Zhan and H. Ding, "Application of contrast-enhanced ultrasound for evaluation of thyroid nodules," *Ultrasonography*, vol. 37, no. 4, p. 288, Oct. 2018.
- [10] S. Kondo *et al.*, "Computer-aided diagnosis of focal liver lesions using contrast-enhanced ultrasonography with perflubutane microbubbles," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1427–1437, Jul. 2017.
- [11] Y. Feng *et al.*, "A deep learning approach for targeted contrast-enhanced ultrasound based prostate cancer detection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 6, pp. 1794–1801, Nov. 2019.
- [12] G. Rizzo *et al.*, "Bayesian quantification of contrast-enhanced ultrasound images with adaptive inclusion of an irreversible component," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 1027–1036, Apr. 2017.
- [13] A. Ignee, M. Jedrejczyk, G. Schuessler, W. Jakubowski, and C. F. Dietrich, "Quantitative contrast enhanced ultrasound of the liver for time intensity curves—Reliability and potential sources of errors," *Eur. J. Radiol.*, vol. 73, no. 1, pp. 153–158, 2010.
- [14] J. Zhang, M. Ding, F. Meng, and X. Zhang, "Quantitative evaluation of two-factor analysis applied to hepatic perfusion study using contrast-enhanced ultrasound," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 259–267, Feb. 2013.
- [15] G. J. Lueck, T. K. Kim, P. N. Burns, and A. L. Martel, "Hepatic perfusion imaging using factor analysis of contrast enhanced ultrasound," *IEEE Trans. Med. Imag.*, vol. 27, no. 10, pp. 1449–1457, Oct. 2008.
- [16] L. Guo *et al.*, "CEUS-based classification of liver tumors with deep canonical correlation analysis and multi-kernel learning," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1748–1751.
- [17] L.-H. Guo *et al.*, "A two-stage multi-view learning framework based computer-aided diagnosis of liver tumors with contrast enhanced ultrasound images," *Clin. Hemorheol. Microcirculat.*, vol. 69, no. 3, pp. 343–354, Jun. 2018.
- [18] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.
- [19] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, Jul. 2015, pp. 2048–2057.
- [20] X. S. Wei, C. L. Zhang, L. Liu, C. Shen, and J. Wu, "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *Proc. IEEE Conf. Asian Conf. Comput. Vis. (ACCV)*, Dec. 2018, pp. 575–591.
- [21] X. Liang, L. Lin, Q. Cao, R. Huang, and Y. Wang, "Recognizing focal liver lesions in CEUS with dynamically trained latent structured models," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 713–727, Mar. 2016.
- [22] N. G. Rognin *et al.*, "Parametric imaging for characterizing focal liver lesions in contrast-enhanced ultrasound," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 57, no. 11, pp. 2503–2511, Nov. 2010.
- [23] Y. Feng, X.-C. Qin, Y. Luo, Y.-Z. Li, and X. Zhou, "Efficacy of contrast-enhanced ultrasound washout rate in predicting hepatocellular carcinoma differentiation," *Ultrasound Med. Biol.*, vol. 41, no. 6, pp. 1553–1560, Jun. 2015.
- [24] C. Huang-Wei *et al.*, "Differential diagnosis of focal nodular hyperplasia with quantitative parametric analysis in contrast-enhanced sonography," *Investigative Radiol.*, vol. 41, no. 3, pp. 363–368, Mar. 2006.
- [25] S. Turco *et al.*, "Contrast-enhanced ultrasound quantification: From kinetic modeling to machine learning," *Ultrasound Med. Biol.*, vol. 46, no. 3, pp. 518–543, Mar. 2020.
- [26] K. Wu, X. Chen, and M. Ding, "Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound," *Optik*, vol. 125, no. 15, pp. 4057–4063, Aug. 2014.
- [27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
- [28] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
- [29] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6450–6459.
- [30] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9910, Oct. 2016, pp. 868–884.
- [31] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.



- [32] M. Wang, C. Lian, D. Yao, D. Zhang, M. Liu, and D. Shen, "Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 8, pp. 2241–2252, Aug. 2020.
- [33] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, and Q. Tian, "A multi-scale spatial-temporal attention model for person re-identification in videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3365–3373, 2020.
- [34] H. Yang, C. Yuan, L. Zhang, Y. Sun, W. Hu, and S. J. Maybank, "STA-CNN: Convolutional spatial-temporal attention learning for action recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 5783–5793, 2020.
- [35] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
- [36] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11205, Sep. 2018, pp. 831–846.
- [37] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11209, Sep. 2018, pp. 413–431.
- [38] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 284–293.
- [39] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.
- [40] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
- [41] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 588–597.
- [42] D. Zhang, X. Dai, and Y. F. Wang, "Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, vol. 11364, Dec. 2018, pp. 712–728.
- [43] P. Wan *et al.*, "CEUS-Net: Lesion segmentation in dynamic contrast-enhanced ultrasound with feature-reweighted attention mechanism," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1816–1819.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
- [45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [46] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv:1607.06450. [Online]. Available: <https://arxiv.org/abs/1607.06450>
- [47] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4743–4752.
- [48] S. Bakas, M. Doulgerakis-Kontoudis, G. J. A. Hunter, P. S. Sidhu, D. Makris, and K. Chatzimichail, "Evaluation of indirect methods for motion compensation in 2-D focal liver lesion contrast-enhanced ultrasound (CEUS) imaging," *Ultrasound Med. Biol.*, vol. 45, no. 6, pp. 1380–1396, Jun. 2019.
- [49] M. Ambai and Y. Yoshida, "CARD: Compact and real-time descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 97–104.
- [50] C. V. Sheena and N. K. Narayanan, "Key-frame extraction by analysis of histograms of video frames using statistical methods," *Procedia Comput. Sci.*, vol. 70, pp. 36–40, Jan. 2015.
- [51] G. B. Cavallari, L. S. F. Ribeiro, and M. A. Ponti, "Unsupervised representation learning using convolutional and stacked auto-encoders: A domain and cross-domain feature space analysis," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 440–446.
- [52] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.
- [53] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.
- [54] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," 2017, arXiv:1705.02953. [Online]. Available: <https://arxiv.org/abs/1705.02953>
- [55] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
- [56] B. Zhang *et al.*, "Utility of contrast-enhanced ultrasound for evaluation of thyroid nodules," *Thyroid*, vol. 20, no. 1, pp. 51–57, Jan. 2010.
- [57] X. Zhou *et al.*, "Diagnostic efficiency of quantitative contrast-enhanced ultrasound indicators for discriminating benign from malignant solid thyroid nodules," *J. Ultrasound Med.*, vol. 37, no. 2, pp. 425–437, Feb. 2018.
- [58] H. Zhao *et al.*, "Diagnostic performance of thyroid imaging reporting and data system (TI-RADS) alone and in combination with contrast-enhanced ultrasonography for the characterization of thyroid nodules," *Clin. Hemorheol. Microcirculat.*, vol. 72, no. 1, pp. 95–106, Jul. 2019.
- [59] Y. He, X. Y. Wang, Q. Hu, X. X. Chen, B. Ling, and H. M. Wei, "Value of contrast-enhanced ultrasound and acoustic radiation force impulse imaging for the differential diagnosis of benign and malignant thyroid nodules," *Frontiers Pharmacol.*, vol. 9, p. 1363, Nov. 2018.
- [60] U. R. Acharya, S. S. Vinitha, F. Molinari, R. Garberoglio, A. Witkowska, and J. S. Suri, "Automated benign & malignant thyroid lesion characterization and classification in 3D contrast-enhanced ultrasound," in *Proc. Eng. Med. Biol. Soc. (EMBC)*, Aug. 2012, pp. 452–455.
- [61] X. Xi *et al.*, "Differentiation of thyroid nodules difficult to diagnose with contrast-enhanced ultrasonography and real-time elastography," *Frontiers Oncol.*, vol. 10, p. 112, Feb. 2020.
- [62] W. Luo *et al.*, "Differential diagnosis of thyroid nodules through a combination of multiple ultrasonography techniques: A decision-tree model," *Exp. Therapeutic Med.*, vol. 19, no. 6, pp. 3675–3683, Mar. 2020.