Thyroid Nodule Diagnosis in Dynamic Contrast-Enhanced Ultrasound via Microvessel Infiltration Awareness

Haojie Han[1], Hongen Liao[2], Daoqiang Zhang[1], Wentao Kong[3], and Fang Chen[1](✉)

[1] Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
chenfang@nuaa.edu.cn

[2] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 10084, China

[3] Department of Ultrasound, Affiliated Drum Tower Hospital, Nanjing University Medical School, Nanjing 21008, China

Abstract. Dynamic contrast-enhanced ultrasound (CEUS) video with microbubble contrast agents reflects the microvessel distribution and dynamic microvessel perfusion, and may provide more discriminative information than conventional gray ultrasound (US). Thus, CEUS video has vital clinical value in differentiating between malignant and benign thyroid nodules. In particular, the CEUS video can show numerous neovascularisations around the nodule, which constantly infiltrate the surrounding tissues. Although the infiltrative of microvessel is ambiguous on CEUS video, it causes the tumor size and margin to be larger on CEUS video than on conventional gray US and may promote the diagnosis of thyroid nodules. In this paper, we propose a novel framework to diagnose thyroid nodules based on dynamic CEUS video by considering microvessel infiltration and via segmented confidence mapping assists diagnosis. Specifically, the Temporal Projection Attention (TPA) is proposed to complement and interact with the semantic information of microvessel perfusion from the time dimension of dynamic CEUS. In addition, we employ a group of confidence maps with a series of flexible Sigmoid Alpha Functions (SAF) to aware and describe the infiltrative area of microvessel for enhancing diagnosis. The experimental results on clinical CEUS video data indicate that our approach can attain an diagnostic accuracy of 88.79% for thyroid nodule and perform better than conventional methods. In addition, we also achieve an optimal dice of 85.54% compared to other classical segmentation methods. Therefore, consideration of dynamic microvessel perfusion and infiltrative expansion is helpful for CEUS-based diagnosis and segmentation of thyroid nodules. The datasets and codes will be available.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_17.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 169–179, 2023.
https://doi.org/10.1007/978-3-031-43987-2_17

Keywords: Contrast-enhanced ultrasound · Thyroid nodule · Dynamic perfusion · Infiltration Awareness

1 Introduction

Contrast-enhanced ultrasound (CEUS) as a modality of functional imaging has the ability to assess the intensity of vascular perfusion and haemodynamics in the thyroid nodule, thus considered a valuable new approach in the determination of benign vs. malignant nodules [1]. In practice, CEUS video allows the dynamic observation of microvascular perfusion through intravenous injection of contrast agents. According to clinical experience, for thyroid nodules diagnosis, there are two characteristic that are important when analyzing CEUS video. 1) Dynamic microvessel perfusion. As shown in Fig. 1(A), clinically acquired CEUS records the dynamic relative intensity changes (microvessel perfusion pattern) throughout the whole examination [2]. 2) Infiltrative expansion of microvessel. Many microvessels around nodules are constantly infiltrating and growing into the surrounding tissue. As shown in Fig. 1(B), based on the difference in lesion size displayed by the two modalities, clinical practice shows that gray US underestimates the size of lesions, and CEUS video overestimates the size of some lesions [3]. Although the radiologist's cognition of microvascular invasive expansion is fuzzy, they think it may promote diagnosing thyroid nodules [1].

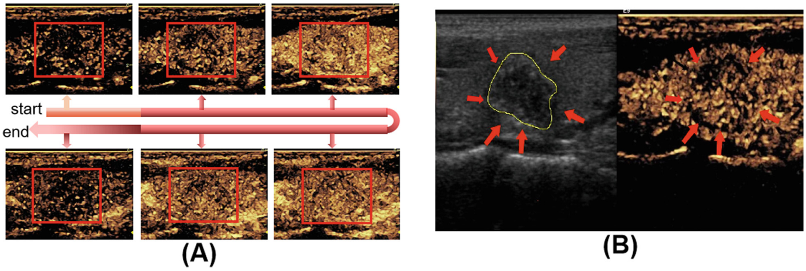

Fig. 1. (A) The dynamic cropped frames in CEUS video. Radiologists identify the area of lesions by comparing various frames, but each individual frame does not effectively describe the lesions area with precision. From the start to the end of the timeline, the three colours represent the change in intensity of the CEUS video over three different time periods. The red border represents the area of the lesion. (B) The size of thyroid nodules described on CEUS is significantly larger than detected through gray US. The yellow line indicates the lesion area labeled by radiologists on gray US, while the red arrow corresponds to the infiltration area or continuous lesion expansion on CEUS. (Color figure online)

Currently, CEUS-based lesion diagnosis methods mainly use the convolution neural network (CNN) to extract spatial-temporal features from dynamic CEUS. Wan et al. [4] proposed a hierarchical temporal attention network which

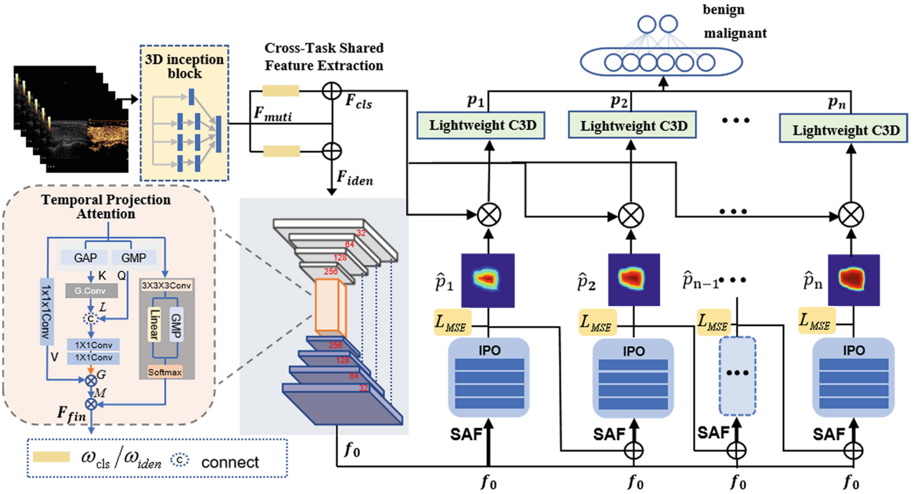

Fig. 2. The overall architecture which contains four parts: cross-task shared feature extraction, temporal-based lesions area recognition, microvessel infiltration awareness and thyroid nodules diagnosis. Here,$\hat{P} = \{\hat{p}_1, \hat{p}_2 \ldots \hat{p}_n\}$ represents the infiltration process of microvessels from gray US to CEUS, and $\hat{p}_n$ is the final segmentation result.

used the spatial feature enhancement for disease diagnosis based on dynamic CEUS. Furthermore, by combing the US modality, Chen et al. [5] proposed a domain-knowledge-guided temporal attention module for breast cancer diagnosis. However, due to artifacts in CEUS, SOTA classification methods often fail to learn regions where thyroid nodules are prominent (As in Appendix Fig. A1) [6]. Even the SOTA segmentation methods cannot accurately identify the lesion area for blurred lesion boundaries, thus, the existing automatic diagnosis network using CEUS still requires manual labeling of pixel-level labels which will lose key information around the tissues [7]. In particular, few studies have developed the CEUS video based diagnostic model inspired by the dynamic microvessel perfusion, or these existing methods generally ignore the influence of microvessel infiltrative expansion. Whether the awareness of infiltrative area information can be helpful in the improvement of diagnostic accuracy is still unexplored.

Here, we propose an explanatory framework for the diagnosis of thyroid nodules based on dynamic CEUS video, which considers the dynamic perfusion characteristics and the amplification of the lesion region caused by microvessel infiltration. Our contributions are twofolds. First, the Temporal Projection Attention (TPA) is proposed to complement and interact with the semantic information of microvessel perfusion from the time dimension. Second, we adopt a group of confidence maps instead of binary masks to perceive the infiltrative expansion area from gray US to CEUS of microvessels for improving diagnosis.

2 Method

The architecture of the proposed framework is shown in Fig. 2. The tasks of lesion area recognition and differential diagnosis are pixel-level and image-level classifications, and some low-level features of these two tasks can be shared interactively [8]. We first fed the CEUS video $I \in \mathbf{R}^{C\times T\times H\times W}$ into the cross-task feature extraction (CFA) module to jointly generate the features F_{iden} and F_{cls} for lesion area recognition and differential diagnosis, respectively. After that, in the temporal-based lesions area recognition (TLAR) module, an enhanced V-Net with the TPA is implemented to identify the relatively clear lesion area which are visible on both gray US and CEUS video. Because microvessel invasion expansion causes the tumor size and margin depicted by CEUS video to be larger than that of gray US, we further adopt a group of confidence maps based on Sigmoid Alpha Functions (SAF) to aware the infiltrative area of microvessels for improving diagnosis. Finally, the confidence maps are fused with F_{cls} and fed into a diagnosis subnetwork based on lightweight C3D [9] to predict the probability of benign and malignant. In the CFA, we first use the 3D inception block to extract multi-scale features F_{muti}. The 3D inception block has 4 branches with cascaded 3D convolutions. Multiple receptive fields are obtained through different branches, and then group normalization and ReLU activation are performed to obtain multi-scale features F_{muti}. Then, we use the cross-task feature adaptive unit to generate the features F_{iden} and F_{cls} required for lesions area recognition and thyroid nodules diagnosis via the following formula [10]:

$$[F_{iden}, F_{cls}] = [\omega_{iden}, \omega_{cls}] * F_{muti} + [F_{muti}, F_{muti}] \tag{1}$$

where $\omega_{iden}, \omega_{cls}$ are the learnable weights.

2.1 Temporal-Based Lesions Area Recognition (TLAR)

The great challenge of automatic recognition of lesion area from CEUS video is that the semantic information of the lesion area is different in the CEUS video of the different microvessel perfusion periods. Especially in the perfusion period and the regression period, the semantic information of lesions cannot be fully depicted in an isolated CEUS frame. Thus, the interactive fusion of semantic information of the whole microvessel perfusion period will promote the identification of the lesion area, and we design the Temporal Projection Attention (TPA) to realize this idea. We use V-Net as the backbone, which consists of four encoder/decoder blocks for TLAR, and the TPA is used in the bottleneck of the V-Net.

Temporal Projection Attention (TPA). Given a feature $F_{4th} \in \mathbf{R}^{C\times T\times \frac{H}{16}\times \frac{W}{16}}$ after four down-sampling operations in encoder, its original 3D feature map is projected [11] to 2D plane to get keys and queries: $K, Q \in \mathbf{R}^{C\times \frac{H}{16}\times \frac{W}{16}}$, and we use global average pooling (GAP) and global maximum pooling (GMP) as temporal projection operations. Here, $V \in \mathbf{R}^{C\times T\times \frac{H}{16}\times \frac{W}{16}}$ is obtained by a single convolution. This operation can also filter out the irrelevant background and display the key information of the lesions. After the temporal projection, a group convolution with a group size of 4 is employed on K to extract the local temporal attention $L \in \mathbf{R}^{C \times \frac{H}{16} \times \frac{W}{16}}$. Then, we concatenate L with Q to further obtain the global attention $G \in \mathbf{R}^{C \times 1 \times \frac{H}{16} \times \frac{W}{16}}$ by two consecutive 1×1 2D convolutions and dimension expend. Those operations are described as follows:

$$K = Q = GAP(F_{4th}) + GMP(F_{4th}) \tag{2}$$

$$G = Expend(Conv(\sigma(Conv(\sigma(Gonv(K)) \oplus Q)))) \tag{3}$$

where $Gonv(\cdot)$ is the group convolution, σ denotes the normalization, "$\oplus$" is the concatenation operation. The global attention G encodes not only the contextual information within isolated query-key pairs but also the attention inside the keys [12]. After that, based on the 2D global attention G, we multiply V and G to calculate the global temporal fusion attention map $M \in \mathbf{R}^{C \times T \times \frac{H}{16} \times \frac{W}{16}}$ to enhance the feature representation.

Meanwhile, to make better use of the channel information, we use $3 \times 3 \times 3$ $Conv$ to get enhanced channel feature $F''_{4th} \in \mathbf{R}^{C_1 \times T \times \frac{H}{16} \times \frac{W}{16}}$. Then, we use parallel average pooling and full connection operation to reweight the channel information of F''_{4th} to obtain the reweighted feature $F'_{4th} \in \mathbf{R}^{C \times T \times \frac{H}{16} \times \frac{W}{16}}$. The obtained global temporal fusion attention maps M are fused with the reweighted feature F'_{4th} to get the output features F_{fin}. Finally, we input F_{fin} into the decoder of the TLAR to acquire the feature map of lesion.

2.2 Microvessel Infiltration Awareness (MIA)

We design a MIA module to learn the infiltrative areas of microvessel. The tumors and margin depicted by CEUS may be larger than those depicted by gray US because of continuous infiltrative expansion. Inspired by the continuous infiltrative expansion, a series of flexible Sigmoid Alpha Functions (SAF) simulate the infiltrative expansion of microvessels by establishing the distance maps from the pixel to lesion boundary. Here, the distance maps [13] are denoted as the initial probability distribution P_D. Then, we utilize Iterative Probabilistic Optimization (IPO) unit to produce a set of optimized probability maps $\hat{P} = \{\hat{p}_1, \hat{p}_2 \ldots \hat{p}_n\}$ to aware the microvessel infiltration for thyroid nodules diagnosis. Based on SAF and IPO, CEUS-based diagnosis of thyroid nodules can make full use of the ambiguous information caused by microvessel infiltration.

Sigmoid Alpha Function (SAF). It is generally believed that the differentiation between benign and malignant thyroid nodules is related to the pixels around the boundaries of the lesion [14], especially in the infiltrative areas of microvessel [3]. Therefore, we firstly build the initial probability distribution P_D based on the distance between the pixels and the annotation boundaries by using SAF in order to aware the infiltrative areas. Here, SAF is defined as follows:

$$SAF_{(i,j)} = C * \left(\frac{2}{1 + e^{\frac{-\alpha D(i,j)}{max(D(i,j))}}} - 1 \right) \tag{4}$$

$$C = \left(1 + e^{-\alpha}\right) / \left(1 - e^{-\alpha}\right); \quad \alpha \in (0, +\infty) \tag{5}$$

where α is the conversion factor for generating initial probability distribution P_D (when $\alpha \rightarrow \infty$, the generated P_D is binary mask); C is used to control the function value within the range of $[0, 1]$; (i, j) is the coordinate point in feature map; $D\,(i, j)$ indicates the shortest distance from (i, j) to lesion's boundaries.

Iterative Probabilistic Optimization (IPO) Unit. Based on the fact that IncepText [15] has experimentally demonstrated that asymmetric convolution can effectively solve the problem of highly variable size and aspect ratio, we use asymmetric convolution in the IPO unit. Asymmetric convolution-based IPO unit can optimize the initial distribution P_D to generate optimized probability maps $\hat{P}$ that can reflect the confidence of benign and malignant diagnosis. Specifically, with the IPO, our network can make full use of the prediction information in low-level iteration layer, which may improve the prediction accuracy of high-level iteration layer. In addition, the parameters in the high-level iteration layer can be optimized through the back-propagation gradient from the high-level iteration layer. IPO unit can be shown as the following formula:

$$\hat{p}_1 = ConvBlock(SAF(f_0, \alpha_1)) \tag{6}$$

$$\hat{p}_i = ConvBlock(SAF((f_0 \oplus \hat{p}_{i-1}), \alpha_i)) \quad i \in (1, n] \tag{7}$$

where "$\oplus$" represents the concatenation operation;*ConvBlock* consists of a group of asymmetric convolutions (e.g., Conv1 × 5, Conv5 × 1 and Conv1 × 1); n denotes the number of the layers of IPO unit. With the lesion's feature map f_0 from the TLAR module, the initial distribution P_D obtained by SAF is fed into the first optimize layer of IPO unit to produce the first optimized probability map $\hat{p}_1$. Then, $\hat{p}_1$ is contacted with f_0, and used to generate optimized probability map $\hat{p}_2$ through the continuous operation based on SAF and the second optimize layer of IPO unit. The optimized probability map $\hat{p}_{i-1}$ provides prior information for producing the next probability map $\hat{p}_i$. In this way, we can get a group of probability map $\hat{P}$ to aware the microvascular infiltration.

2.3 Loss Function

With continuous probability map $\hat{P}$ obtained from MIA, $\hat{P}$ are multiplied with the feature F_{cls}. Then, these maps are fed into a lightweight C3D to predict the probability of benign and malignant, as shown in Fig. 2. We use the mean square error L_{MSE} to constrain the generation of $\hat{P}$. Assuming that the generated $\hat{P}$ is ready to supervise the classification network, we want to ensure that the probability maps can accurately reflect the classification confidence. Thus, we design a task focus loss L_{ta} to generate confidence maps P, as follows:

$$L_{MSE} = \sum_{i=1}^{n} \frac{1}{\varOmega} \sum_{p \in \varOmega} \|\boldsymbol{g}_i(pi), \hat{\boldsymbol{p}}_i(pi)\|_2 \tag{8}$$

$$L_{ta} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|\boldsymbol{p}_i - \widehat{\boldsymbol{p}}_i\|_2^2 + \log\sigma \tag{9}$$

where $\boldsymbol{g}_i$ is the label of $\hat{\boldsymbol{p}}_i$, which is generated by the operation of $SAF(D_{(i,j)}, \alpha_i)$; pi denotes pixel in the image domain Ω, σ is a learnable parameter to eliminate the hidden uncertainty information.

For differentiating malignant and benign, we employ a hybrid loss L_{total} that consists of the cross-entropy loss L_{cls}, the loss of L_{MSE} computing optimized probability maps $\hat{P}$, and task focus loss L_{ta}. The L_{total} is denoted as follows:

$$L_{total} = \lambda_1 \cdot L_{cls} + \lambda_2 \cdot L_{MSE} + \lambda_3 \cdot L_{ta} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the hyper-parameters to balance the corresponding loss. As the weight parameter, we set $\lambda_1, \lambda_2, \lambda_3$ are 0.5,0.2,0.3 in the experiments.

3 Experiments

Dataset. Our dataset contained 282 consecutive patients who underwent thyroid nodule examination at Nanjing Drum Tower Hospital. All patients performed dynamic CEUS examination by an experienced sonographer using an iU22 scanner (Philips Healthcare, Bothell, WA) equipped with a linear transducer L9-3 probe. These 282 cases included 147 malignant nodules and 135 benign nodules. On the one hand, the percutaneous biopsy based pathological examination was implemented to determine the ground-truth of malignant and benign. On the other hand, a sonographer with more than 10 years of experience manually annotated the nodule lesion mask to obtain the pixel-level ground-truth of thyroid nodules segmentation. All data were approved by the Institutional Review Board of Nanjing Drum Tower Hospital, and all patients signed the informed consent before enrollment into the study.

Implementation Details. Our network was implemented using Pytorch framework with the single 12 GB GPU of NVIDIA RTX 3060. During training, we first pre-trained the TALR backbone via dice loss for 30 epochs and used Adam optimizer with learning rate of 0.0001. Then, we loaded the pre-trained weights to train the whole model for 100 epochs and used Adam optimizer with learning rate of 0.0001. Here, we set batch-size to 4 during the entire training process The CEUS consisted the full wash-in and wash-out phases, and the resolution of each frame was (600×800). In addition, we carried out data augmentation, including random rotation and cropping, and we resize the resolution of input

Table 1. Quantitative lesion recognition results are compared with SOTA methods and ablation experiments.

Network	UNet3D [18]	V-Net [16]	TransUNet [17]	V-Net+ TPA	V-Net+ TPA+SAF	V-Net+ TPA+IPO	**Ours**
DICE(%)↑	73.63 ± 5.54	77.94 ± 4.77	72.84 ± 6.88	81.22 ± 4.18	82.96 ± 4.11	83.32 ± 4.03	$\mathbf{85.54 \pm 4.93}$
Recall(%)↑	79.96 ± 4.39	83.17 ± 5.05	77.69 ± 5.18	83.96 ± 3.66	88.71 ± 3.92	89.45 ± 3.77	$\mathbf{90.40 \pm 5.93}$
IOU(%)↑	60.56 ± 5.83	65.20 ± 5.09	59.83 ± 6.19	69.96 ± 3.58	72.63 ± 3.15	73.26 ± 4.03	$\mathbf{74.99 \pm 4.72}$

frames to (224 × 224). We adopted 5-fold cross-validation to achieve quantitative evaluation. Three indexes including Dice, Recall, and IOU, were used to evaluate the lesion recognition task, while five indexes, namely average accuracy (ACC), sensitivity (Se), specificity (Sp), F1-score (F1), and AUC, were used to evaluate the diagnosis task.

Experimental Results. As in Table 1, we compared our method with SOTA method including V-Net, Unet3D, TransUnet. For the task of identifying lesions, the index of Recall is important, because information in irrelevant regions can be discarded, but it will be disastrous to lose any lesion information. V-Net achieved the highest Recall scores compared to others; thus, it was chosen as the backbone of TLAR. Table 1 revealed that the modules (TPA, SAF, and IPO) used in the network greatly improved the segmentation performance compared to baseline, increasing Dice and Recall scores by 7.60% and 7.23%, respectively. For the lesion area recognition task, our method achieved the highest Dice of 85.54% and Recall of 90.40%, and the visualized results were shown in Fig. 3.

Table 2. Quantitative diagnostic results are compared with SOTA methods and ablation experiments.

Network	ACC (%) ↑	Se (%) ↑	Sp (%) ↑	F1 (%) ↑	AUC (%) ↑
C3D+Mask [19]	76.61 ± 4.13	88.05 ± 3.73	76.92 ± 3.17	77.52 ± 3.32	88.05 ± 3.01
R3D+Mask [20]	77.01 ± 3.25	87.17 ± 4.08	79.49 ± 3.88	87.05 ± 3.93	83.10 ± 2.35
R2plus1D+Mask [21]	78.88 ± 2.81	85.02 ± 2.49	77.45 ± 3.43	87.75 ± 2.79	81.10 ± 3.82
ConvLSTM+Mask [22]	78.59 ± 4.44	84.37 ± 3.66	76.26 ± 4.26	80.22 ± 3.92	85.95 ± 3.73
Baseline+Mask	79.29 ± 2.58	89.83 ± 1.80	80.84 ± 3.04	86.01 ± 2.00	88.25 ± 2.84
Baseline+TLAR	81.10 ± 2.24	84.97 ± 1.28	81.58 ± 2.74	82.49 ± 1.81	88.67 ± 1.96
Baseline+TLAR+SAF	84.15 ± 1.78	89.90 ± 0.94	79.08 ± 1.85	83.95 ± 1.79	89.90 ± 1.97
Baseline+TLAR+IPO	86.56 ± 2.45	92.58 ± 2.38	79.93 ± 2.53	86.41 ± 1.36	93.33 ± 2.74
Ours	**88.79 ± 1.40**	**94.26 ± 1.68**	**88.37 ± 1.80**	**90.41 ± 1.85**	**94.54 ± 1.54**

To evaluate the effectiveness of the baseline of lightweight C3D, we compared the results with SOTA video classification methods including C3D, R3D, R2plus1D and ConvLSTM. For fair comparison, all methods used the manually annotated lesion mask to assist the diagnosis. Experimental results in Table 2 revealed that our baseline network could be useful for the diagnosis. With the effective baseline, the introduced modules including TLAR, SAF and IPO further improved the diagnosis accuracy, increasing the accuracy by 9.5%. The awareness of microvascular infiltration using SAF and IPO unit was helpful for CEUS-based diagnosis, as it could improve the diagnosis accuracy by 7.69% (As in Table 2). As in Appendix Fig. A1, although SOTA method fails to focus on lesion areas, our method can pinpoint discriminating lesion areas.

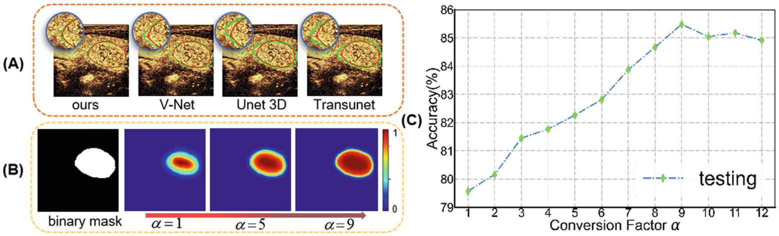

Fig. 3. (A)Comparison of the visualised results with the SOTA method, green and red contours is the automatically recognized area and ground-truth. By enlarging the local details, we show that our model can obtain the optimal result (More visuals provided in Appendix Fig. A2 and Fig. A3.). (B) Microvascular infiltration was simulated from gray US to CEUS via a set of confidence maps. (C) Influence of α values. (Color figure online)

Influence of α Values. The value of α in *SAF* is associated with simulating microvessel infiltration. Figure 3 (C) showed that the diagnosis accuracy increased along with the increment of α and then tended to become stable when α was close to 9. Therefore, for balancing the efficiency and performance, the number of IPO was set as n = 3 and α was set as $\alpha = \{1, 5, 9\}$ to generate a group of confidence maps that can simulate the process of microvessel infiltration. (More details about the setting of n is in Appendix Fig. A4 of the supplementary material.)

4 Conclusion

The microvessel infiltration leads to the observation that the lesions detected on CEUS tend to be larger than those on gray US. Considering the microvessel infiltration, we propose an method for thyroid nodule diagnosis based on CEUS videos. Our model utilizes a set of confidence maps to recreate the lesion expansion process; it effectively captures the ambiguous information caused by microvessel infiltration, thereby improving the accuracy of diagnosis. This method is an attempt to eliminate the inaccuracy of diagnostic task due to the fact that gray US underestimates lesion size and CEUS generally overestimates lesion size. To the best of our knowledge, this is the first attempt to develop an automated diagnostic tool for thyroid nodules that takes into account the effects of microvessel infiltration. The way in which we fully exploit the information in time dimension through TPA also makes the model more clinically explanatory.

References

1. Radzina, M., Ratniece, M., Putrins, D.S., et al.: Performance of contrast-enhanced ultrasound in thyroid nodules: review of current state and future perspectives. Cancers **13**(21), 5469 (2021)

2. Yongfeng, Z., Ping, Z., Hong, P., Wengang, L., Yan, Z.: Superb microvascular imaging compared with contrast-enhanced ultrasound to assess microvessels in thyroid nodules. J. Med. Ultrasonics **47**(2), 287–297 (2020). https://doi.org/10.1007/s10396-020-01011-z
3. Jiang, Y.X., Liu, H., Liu, J.B., et al.: Breast tumor size assessment: comparison of conventional ultrasound and contrast-enhanced ultrasound. Ultrasound Med. Biol. **33**(12), 1873–1881 (2007)
4. Wan, P., Chen, F., Zhang, D., et al.: Hierarchical temporal attention network for thyroid nodule recognition using dynamic CEUS imaging. IEEE Trans. Med. Imaging **40**(6), 1646–1660 (2021)
5. Chen, C., Wang, Y., et al.: Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Trans. Med. Imaging **40**(9), 2439–2451 (2021)
6. Manh, V. T., Zhou, J., Jia, X., Lin, Z., et al.: Multi-attribute attention network for interpretable diagnosis of thyroid nodules in ultrasound images. IEEE Trans. Ultrason. Ferroelect. Frequency Control 69(9), 2611–2620 (2022)
7. Moon, W.K., Lee, Y.W., et al.: Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images. Comput. Meth. Programs Biomed. **146**, 143–150 (2017)
8. Golts, A., Livneh, I., Zohar, Y., et al.: Simultaneous detection and classification of partially and weakly supervised cells. In: Karlinsky, L., et al. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 313–329. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_16
9. Wang, Y., Li, Z., Cui, X., Zhang, L., et al.: Key-frame guided network for thyroid nodule recognition using ultrasound videos. In: Wang, L., et al. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 238–247. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_23
10. Wang, X., Jiang, L., Li, L., et al.: Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis. IEEE Trans. Med. Imaging **40**(9), 2463–2476 (2021)
11. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C.: Axial-DeepLab: stand-alone axial-attention for panoptic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 108–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_7
12. Jiang, Y., Zhang, Z., Qin, S., et al.: APAUNet: axis projection attention UNet for small target in 3D medical segmentation. In: Proceedings of the Asian Conference on Computer Vision, pp. 283–298 (2022)
13. Zhang, S., Zhu, X., Chen, L., Hou, J., et al.: Arbitrary shape text detection via segmentation with probability maps. IEEE Trans. Pattern Anal. Mach. Intell. **14** (2022). https://doi.org/10.1109/TPAMI.2022.3176122
14. Gómez-Flores, W., et al.: Assessment of the invariance and discriminant power of morphological features under geometric transformations for breast tumor classification. Comput. Methods Programs Biomed. **185**, 105173 (2020)
15. Yang, Q., et al.: Inceptext: a new inception-text module with deformable psroipooling for multi-oriented scene text detection. In: IJCAI, pp. 1071–1077(2018)
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: IEEE Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
17. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Tran, D., et al.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
20. Tran, D., Wang, H., et al.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
21. Tran, D., Wang, H., et al.: A closer look at spatiotemporal convolutions for action recognition . In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
22. Mutegeki, R., Han, D. S., et al.: A CNN-LSTM approach to human activity recognition. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 362–366 (2022)