Anatomical-Functional Fusion Network for Lesion Segmentation Using Dual-View CEUS

Peng Wan[1], Chunrui Liu[2], and Daoqiang Zhang[1](✉)

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

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

Abstract. Dual-view contrast-enhanced ultrasound (CEUS) has been widely applied in lesion detection and characterization due to the provided anatomical and functional information of lesions. Accurate delineation of lesion contour is important to assess lesion morphology and perfusion dynamics. Although the last decade has witnessed the unprecedented progress of deep learning methods in 2D ultrasound imaging segmentation, there are few attempts to discriminate tissue perfusion discrepancy using dynamic CEUS imaging. Combined with the side-by-side gray-scale US view, we propose a novel anatomical-functional fusion network (AFF-Net) to fuse complementary imaging characteristics from dual-view dynamic CEUS imaging. Towards a comprehensive characterization of lesions, our method mainly tackles with two challenges: 1) how to effectively represent and aggregate enhancement features of the dynamic CEUS view; 2) how to efficiently fuse them with the morphology features of the US view. Correspondingly, we design the channel-wise perfusion (PE) gate and anatomical-functional fusion (AFF) module with the goal to exploit dynamic blood flow characteristics and perform layer-level fusion of the two modalities, respectively. The effectiveness of the AFF-Net method on lesion segmentation is validated on our collected thyroid nodule dataset with superior performance compared with existing methods.

Keywords: Multi-modality Fusion · Nodule Segmentation · Contrast-enhanced ultrasound · Co-attention

1 Introduction

Ultrasound (US), as the first-line diagnostic tool in early screening and diagnosis, has become increasingly important in clinical assessment due to the advantages of cost-effectiveness, portability, non-ionizing radiation, and real-time assessment. Thyroid nodule are a common finding in the general population with a

P. Wan and C. Liu—Contributed equally to this work.

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 244–256, 2023.
https://doi.org/10.1007/978-3-031-46664-9_17

detection rate of 50% to 60% [1,2]. Ultrasonic features like nodule size, location, shape regularity, margin smoothness, and extra-thyroidal extension are important imaging findings for malignancy risk prediction [3], postoperative assessment [5], and fine-needle aspiration biopsy planning [4]. Thus, accurate nodule segmentation is an indispensable step in clinical practice. In addition to traditional anatomical imaging (B-mode ultrasound, BUS), the emerging functional imaging (contrast-enhanced ultrasound, CEUS) allows for a real-time observation of microvascular perfusion within thyroid gland by enhancing blood flow signals from small vessels [6,7]. Generally, radiologists perform a comprehensive analysis of morphology features in gray-scale US and perfusion features in contrast-enhanced US, but this step requires a high level of expertise and is susceptible to subjective errors.

Although several machine learning or deep learning techniques have been proposed for segmenting thyroid nodules using US imaging, including active contours [10], fuzzy clustering [9], and fully convolution network [8,11–13] etc., segmentation performances of these methods are still limited. One major limitation is that these methods have not fully exploited ultrasonic characteristics complementarity in the segmentation task. Taking cystic nodules as example, gray-scale US is more sensitive to internal hypoechoic regions. Nevertheless, due to the infiltrative growth pattern, we might observe a vague or incomplete boundary since marginal echoic intensity differences become much smaller. In case of that, contrast-enhanced US could complement this by highlighting the varying hemodynamic changes around marginal regions, assisting nodule localization and boundary delineation. Another limitation is that existing CEUS based segmentation methods depend on a preselected a reference frame with relatively distinguished contours, ignoring dynamic blood perfusion information. Actually, perfusion discrepancy might consists in initial enhancement, progression to ultimate wash out. Therefore, it is necessary to reason over the whole perfusion process to sufficiently mine enhancement discrepancy between nodule and thyroid gland.

From the perspective of multi-modality imaging segmentation, it is of great importance to exploit the complementarity of different sort of imaging. Towards this goal, Dolz et al. [39] extend the definition of dense connectivity to multimodal streams, such that dense connectivity within each stream and across different streams could enhances the modality information flow while facilitates the network training. On the other hand, attention mechanism also arouse considerable interest in exploiting inter-dependencies of different modalities, instead of simple summation or concatenation operation. Chen et al. [40] proposes a 3D convolutional block to produce the spatial map highlighting relevant image regions from multiple sources. According to imaging prior knowledge, one MR modality is picked as the master modality and the other is treated as an assistant modality. Information fusion is conducted by transferring the attention map learned from the master stream (teacher network) to supervise the training of the assistant stream (student network). Intuitive, sufficient inter-modality interaction at different-level feature abstraction could ensure enough freedom to capture

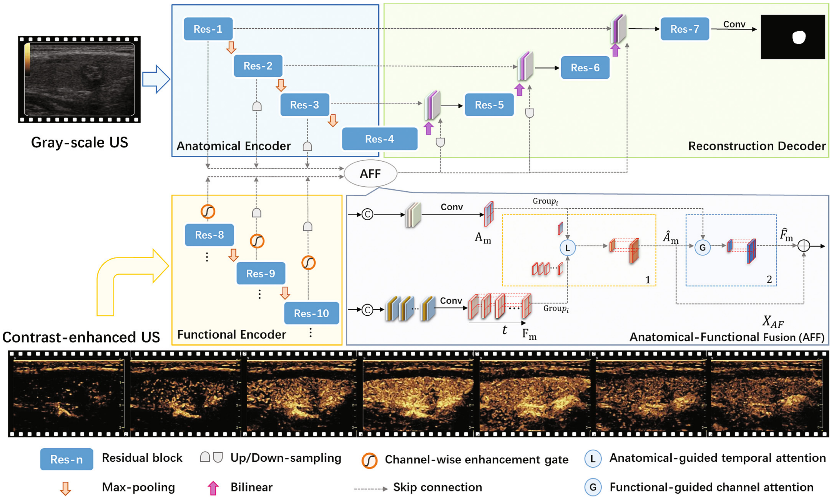

Fig. 1. Illustration of the proposed Anatomical-Functional Fusion Network (AFF-Net).

complex dependencies between modalities. Nonetheless, the optimal layer-level fusion method would vary with specific modalities, leaving an open question for our dual-view CEUS segmentation problem.

In this paper, we propose an anatomical-functional fusion network (AFF-Net) for thyroid nodule segmentation using dual-screen CEUS imaging. For simplicity, we term the morphological and echoic characteristics in gray-scale US view as anatomical features, and dynamic enhancement patterns depicting the real-time blood supply in contrast-enhanced US view as functional features. Figure 1 shows a schematic diagram of our AFF-Net model, which consists of modality-specific encoders and reconstruction decoder, as well as the specifically designed anatomical-functional fusion (AFF) module. By sequentially attending to feature representations of dynamic enhancement patterns and static morphological features, the introduced co-attention mechanism in AFF module integrates multiple US modalities in a layer-level fusion manner. To fully exploit enhancement characteristics, we also introduce a channel-wise enhancement (CE) gate to jointly model enhancement appearances at single point and intensity variations among adjacent frames. We validate the model performance on our collected dual-view thyroid dataset.

2 Related Work

2.1 Medical Imaging Segmentation

For almost a decade, deep learning methods represented by fully convolutional networks (FCN) have pushed medical imaging segmentation into a considerable maturity level both in accuracy and robustness [21,22]. Characterized by a U-shape encoder-decoder architecture, FCN has becomes the basic architecture in various medical segmentation tasks, including cardiac MRI [23], thyroid US [24] and abdominal CT [25]. The former encoder is responsible for representation learning by enhancing pixel-wise discrimination ability, while the latter decoder is the founding part to fuse features from multiple encoding hierarchies. To be aware of different-scale objects, a series of multi-scale representation learning strategies have been proposed, including Gaussian (Laplacian) image pyramid [26,27], atrous spatial pyramid pooling [28], dilated convolution [29], and pyramidal convolution [30]. As for the global context modeling, global context network (GCNet) combines a simplified self-attention mechanism and squeeze-excitation mechanism. As for the basic convolution operation, SEgmentation TRansformer (SETR) [32] replaces it with a pure transformer structure, which also achieves competitive performance.

2.2 Multi-modality Imaging Segmentation

Multi-modal medical imaging (e.g., CT, PET, MRI and US, et al.) has achieved extensive application in comprehensive characterization of morphological, pathophysiological and molecular features of tumors. To exploit the complementarity of different sort of imaging, an increasing number of deep multi-modal methods have emerged recently [33–35]. As mentioned above, feature fusion can be realized at three stages. Among which, early-fusion refers to stacking raw imaging or low-level features channel-wisely by assuming a linear inter-modality relationship [36]. Actually, imaging characteristics from distinct modalities are heterogeneous more than complementary since the imaging acquisition processes differ greatly from each other. To model inter-modality correlation in a higher level feature space, the rest two fusion strategies adopt a multi-path network structure so as to extract a hierarchical representation separately using the state-of-the-art design of each modality. For late fusion, high-level feature maps from different paths are fused only at the stage of model prediction. To facilitate knowledge transfer among different streams (modalities), information fusion is performed in a hierarchical way in the layer-level fusion. As suggested in studies [37] [38], layer-level fusion has the potential to be the optimal fusion way.

3 Materials and Method

3.1 Dataset

In this study, we totally collected 114 dual-screen CEUS videos from patients who attended xx Hospital for thyroid ultrasound examination. All examinations were performed on a Philips iU22 scanner (Philips Medical Systems, Best, the Netherlands) at a low mechanical index ≤ 0.12 using the second-generation contrast agents SonoVue (Bracco SpA, Milan, Italy). Dual-screen CEUS videos were exported as AVI video files with the spatial resolution 600×800. Each video has a duration at least 3 min with a framerate of 15 fps, recording the complete thyroid perfusion process. Each examination was performed by an expert with over 10-year clinical experience, and annotated by at least two senior radiologists to reduce inter-observer variabilities. Each radiologist first reviewed the whole CEUS video, and then selected an optimal frame to contour the boundary. Approval was obtained by the ethics review board of local hospital and the informed consent was obtained from patients before this study.

3.2 Anatomical-Functional Fusion Network (AFF-Net)

Architecture. As illustrated in Fig. 1, we adopt a two-stream U-shape structure to construct our AFF-Net model. In the *encoding* phase, the backbone of Anatomical Encoder consists of four residual blocks separated by 2×2 max-pooling layer. Each block has two 3×3 Conv layers (all with unit stride and zero-padding), followed by the batch normalization and ReLU activation. Each layer is connected to the input of the previous layer. The number of channels is $[16, 16; 32, 32; 64, 64; 64, 64]$. As for Functional Encoder, we adopt three stacked residual blocks with the channel number $[16, 16; 32, 32; 64, 64]$. Besides, we introduce the channel-wise enhancement (CE) gate to explicitly represent inter-frame intensity variations. In the *decoding* phase, anatomical-functional fusion (AFF) module is used to fuse dual-modal feature maps from multiple encoding scales. Along the up-sampling path, to-be-fused ultrasonic representations comprise three components, 1) up-sampled anatomical map generated by the deconvolution layer; 2) high-resolution anatomical map passed by the skip connection; and 3) down-scaled multi-modal map output by the AFF module. The up-sampling path is composed of three sequential residual blocks (channels: $[64, 64, 32, 32, 16, 16]$) separated by the deconvolution layer. Finally, pixel-wise category map P is reconstructed on the fused multi-modal features using a 1-channel 1×1 Conv layer, normalized by a *sigmoid* layer.

Channel-Wise Enhancement Gate. Given sequential enhancement appearance feature maps $\mathbf{M} \in R^{T \times C \times H \times W}$, where T, C, H, W denote the temporal, channel and two spatial dimensions respectively, we first apply a 1×1 2D convolution to reduce feature channels $\mathbf{M}^r(t) = Conv_r * \mathbf{M}(t)$, r is the reduction factor set to 4. Based on that, feature-level enhancement dynamics $E\,(t)$ is approximately represented as inter-frame feature difference between time step t and $t+1$,

$$\mathbf{E}\,(t) = Conv_c * \mathbf{M}^r\,(t+1) - \mathbf{M}^r(t), t \in [1, T-1] \tag{1}$$

where $\mathbf{E}\,(t) \in R^{C/r \times H \times \mathrm{W}}$ is the enhancement map at time step t, $Conv_c$ is a 3×3 channel-wise convolution. In this way, we could obtain $T-1$ enhancement variations representations. To keep temporal consistency, we append an all-zero enhancement map $\mathbf{E}(T)$ at time step T. Then, sequential variations representation maps are convolved by a 1×1 Conv layer to restore channel dimension to C. Finally, we obtain the combined perfusion representation $\mathbf{F} = \mathbf{M} + \mathbf{E}$ via an element-wise summation between the input enhancement appearance M and the sequential enhancement variation $\mathbf{E} = [\mathbf{E}(1), \mathbf{E}(2), \ldots \mathbf{E}(T)] \in R^{T\times C\times H\times \mathrm{W}}$. The behind intuition is that significant intensity variations of contrast-enhanced US view correlate with the real-time changes of spatial distribution of contrast agents, which is expected to trace enhancement discrepancy between lesion and normal tissues.

3.3 Anatomical-Functional Fusion Module

It is worth noting that conventional gray-scale US and contrast-enhanced US actually reflect the thyroid nodule status by complementarily different views. That is, morphological features in gray-scale US are intrinsically correlated with blood flow features in contrast-enhanced US. Therefore, leveraging the semantic consistency between modalities, alternating co-attention mechanism is adopted in our anatomical-functional fusion (AFF) module, which co-attends to both modalities sequentially to distinguish important components for nodule boundary recognition.

Multi-scale Fusion and Grid Split. Given anatomical (functional) features A^s (F_t^s) from different scales s, we rescale them into a common resolution (equaling to the output of the first residual block) by bilinear interpolation, and merge them along channels, $\mathbf{A}_m = Conv_r\left[A^1; A^2; A^3\right](\mathbf{F}_{m,t} = Conv_r\left[F_t^1; F_t^2; F_t^3\right])$, where r is channel reduction factor set to 16. Considering the spatial correspondences, our AFF module restricts inter-modal interactions within the same region, which is greatly different from co-attention mechanism in Visual Question Answering [14,15,19,20] that builds associations between all pairs of image-question locations. Thus, we split the multi-scale anatomical (functional) map $\mathbf{A}(\mathbf{F})$ into N regular grids to co-attend both modalities.

Anatomical-Guided Temporal Attention. To evaluate which contrast frames should be attended or overlooked, the first step is to generate temporal attention under the anatomical guidance. For each i-th grid, we summarize the anatomical-guide attention operation as $\hat{\mathbf{F}}^i = L\left(\mathbf{F}^i, p^i\right)$, where $i = \{1, 2, \ldots N\}$, $\mathbf{F}^i$ and p^i denote the combined enhancement representation and anatomical feature, respectively. Specifically, global average pooling (GAP) is used to summarize the spatial information of $\mathbf{A}^i$, which is then transformed by a fully-connected layer W_A to generate the anatomical guidance p^i,

$$p^i = W_A * GAP\left(\mathbf{A}^i\right) \tag{2}$$

Based on that, temporal attention score $\mathbf{s}$ is calculated by the dot-product between p^i and the respective enhancement descriptor $f_t^i = GAP\left(\mathbf{F}^i\right)$, aiming at highlighting temporal points with significant appearance or intensity variations.

$$s_t = \sigma\left(\langle p^i, f_t^i\rangle\right) \tag{3}$$

where $\sigma(\cdot)$ denotes the *sigmoid* function for normalization. And thus, attentive enhancement representation $\hat{\mathbf{F}}^i$ is calculated by the weighted sum $\hat{\mathbf{F}}^i = \sum_{t=1}^{T} s_t \cdot \mathbf{F}_t^i$.

Functional-Guided Channel Attention. Apart from identifying salient contrast frames to focus on, we also need to emphasize important channels of gray-scale US map, which are closely associated with essential attributes, such as some kind of edges, low echoes, and boundaries. Similar to [17], we propose the functional-guide channel attention operator in each grid, described as $\hat{\mathbf{A}}^i = G\left(\mathbf{A}^i, q^i\right)$, where q^i is the functional guidance from the global average pooling of $\hat{\mathbf{F}}^i$, $\hat{\mathbf{A}}^i$ is the recalibrated anatomical map.

To generate recalibration signal, we first squeeze the spatial information of A^i into the deep anatomical descriptors a^i using GAP, and then predict a joint representation based on anatomical descriptor a^i and functional guidance q^i as follows,

$$K = W_F\left[a^i; q^i\right] \tag{4}$$

where $K \in R^{C^B}$, $W_F \in R^{C^B \times C^{F+B}}$. Finally, K is normalized by sigmoid layer to recalibrate the anatomical map, producing the recalibration representation $\hat{\mathbf{A}}_i = \mathbf{A}_i \odot \sigma\left(K\right)$, where $\odot$ denotes the channel-wise product operation.

As described above, the alternating anatomical-functional attention mechanism is independently performed in each spatial grouping. Finally, the AFF module outputs the fused representation $\mathbf{X}_{AF}$ by combining attended anatomical and functional features $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{F}}^i$ via element-wise summation. In the decoding stage, $\mathbf{X}_{AF}$ is rescaled to the match the resolution before each residual block.

3.4 Implementation and Loss Function

The proposed AFF-Net was implemented using deep learning framework Pytorch and run on a single GPU (NVIDIA TITAN RTX, 24 GB). Considering temporal redundancy of raw CEUS videos, we adopted a temporal pruning strategy [18] to screen out informative contrast subsequences with the length of $T = 7$. Accordingly, one single gray-scale US image and the accompanying contrast-enhanced US subsequence were fed as two modalities into our AFF-Net, as was common in the baseline and competing methods. Model parameters are updated using the Adam optimizer with the default parameters. The learning rate was initialized to 0.001 and adjusted using cosine annealing schedule every 30 epochs. We used a small batch size of 2 and terminated the learning process when validation performance begins to convergence. Our AFF-Net was trained using the Dice loss,

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \varepsilon} \tag{5}$$

where N is the number of pixels in the image, $p_i \in [0, 1]$ is the predicted probability of i_{th} pixel belonging to the lesion area, $y_i \in \{0, 1\}$.

4 Experiments and Results

Experimental Setup. In our experiments, we adopted the standard setup of 5-fold cross-validation for performance evaluation and comparison of our method

Table 1. Comparison with State-of-the-art Methods and Baselines on the task of thyroid nodule segmentation.

Methods	Fusion	Thyroid nodule			
		DSC(%)	IoU(%)	HD	$p \leq 0.05$
MC-CNN	Conv_0	75.36 ± 2.38	61.17 ± 3.10	9.39 ± 0.51	$\star$
	Conv_1	76.38 ± 2.43	62.53 ± 3.13	8.95 ± 0.41	$\star$
MB-CNN	Average	76.99 ± 2.42	63.09 ± 3.09	8.75 ± 0.28	$\star$
	Majority	76.10 ± 2.08	61.84 ± 2.60	9.151 ± 0.45	$\star$
HyperDenseNet	–	79.76 ± 1.99	$66.54 \pm 2.55^\star$	8.22 ± 0.37	–
Co-learning	–	77.34 ± 2.19	63.76 ± 2.83	9.39 ± 0.47	$\star$
MMTM	–	78.04 ± 2.81	64.26 ± 3.61	9.33 ± 0.41	$\star$
AFF-Net	–	$\mathbf{81.74 \pm 1.73}$	$\mathbf{69.40 \pm 2.18}$	$\mathbf{8.50 \pm 0.36}$	–

* denotes a significant difference compared with our method, the last column denotes significant comparisons for all three metrics.

and competing methods, as well as all baselines. In this paper, segmentation performance was evaluated by three metrics, including Dice Similarity Coefficie (DSC), Intersection over Union (IoU) and Hausdorff distance (HD) [16]. The first three metrics measures the degree of overlap between segmentation result S and ground truth Y, and HD measures boundary distances. For all experimental comparisons, we computed the p-value with the two-sample t-test.

We first compared our AFF-Net method with several fusion baselines. 1) Multi-channel (MC) CNN, implementing multi-modal US fusion via channel-wise concatenation at the network input (Conv_0) or after first convolution block (Conv_1); 2) Multi-channel (MB) CNN, implementing a late fusion of segmentation results by average or majority voting, where each modality was processed separately. Then, we compare it with more complex layer-wise fusion structures, including 1) HyperDenseNet that extends the dense connectivity to a multi-branch structure; 2) Co-learning Network that derives a spatially varying fusion map at each decoding scale; 3) Multimodal transfer module (MMTM) that recalibrates multi-modal tensors along the channel dimension. In our implementation, we replace the original 2D convolution with 3D ones, aiming at learn spatial-temporal features from dynamic contrast-enhanced US view.

Baselines and Competing Methods: Quantitative segmentation results are summarized in Table 1. We observe that the layer-level fusion of deep features from different modalities achieves a superior performance over the manner of early-level and late-level fusion. And our proposed AFF-Net achieves the largest overall improvements, these improvements are statistically significant compared to all baselines, verifying the effectiveness of cross-modality imaging fusion and enhancement dynamics representation in the task of thyroid nodule segmentation. By allowing dense connectivity between encoding streams, HyperDenseNet achieves the smallest mean boundary distance of 8.22, and comparable

Table 2. Comparative results of ablation analysis.

Methods	Thyroid nodule			
	DSC(%)	IoU(%)	HD	$p \leq 0.05$
A-Net	74.22 ± 3.13	59.72 ± 4.03	9.65 ± 0.36	$\star$
F-Net	73.88 ± 2.77	59.40 ± 3.54	9.43 ± 0.44	$\star$
AFF-Net-C	79.08 ± 1.79	$65.70 \pm 2.26^\star$	$8.74 \pm 0.42^\star$	–
AFF-Net	$\mathbf{81.74 \pm 1.73}$	$\mathbf{69.40 \pm 2.18}$	$\mathbf{8.50 \pm 0.36}$	–

* denotes a significant difference compared with our method, the last column denotes significant comparisons for all three metrics.

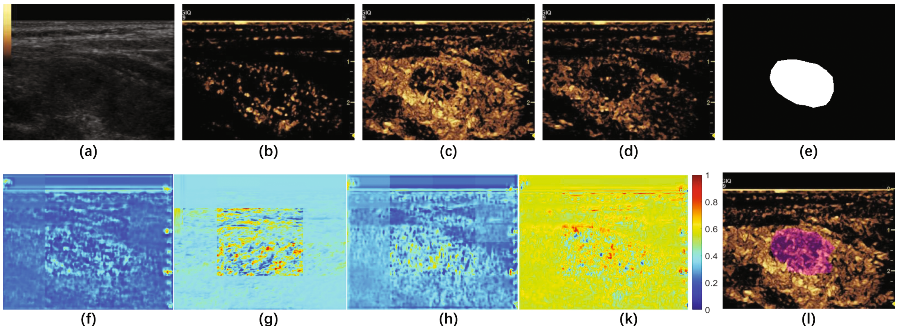

Fig. 2. (a) Gray-scale US; (b–d) Dynamic contrast-enhanced US; (e) Ground-truth; (f) Highlighted anatomical channel $\mathbf{A}_c$; (g)Significant enhancement point $\mathbf{F}_t$; (h) Attend anatomical map $\hat{\mathbf{A}}$; (k) Temporally aggregated functional map $\hat{\mathbf{F}}$; (L)Segmentation result $\mathbf{P}$. For illustration, we normalize the feature values into the range of $[0 - 1]$.

performances in terms of mean DSC 79.76% vs. 81.74% and IoU 66.54% vs. 69.40%, respectively. Another interesting finding is that channel-wise attention in MMTM outperforms spatial-channel-wise attention in Co-learning method. It indicates that deriving a more complex weighting tensor might not be well suitable for feature fusion in the task of nodule segmentation using dual-screen CEUS imaging. For an more intuitive understanding, we provide an visualization of our model in Fig. 2, including the model prediction and the intermediate feature maps generated by the AFF module.

Ablation Analysis: To evaluate the usefulness of multi-modal US fusion and two major components of our method (i.e., CE gate and AFF module), we compare AFF-Net with its three variants, i.e., 1) A-Net, which removes the branch of enhancement features learning from contrast-enhanced US view; 2) F-Net, which removes the branch of morphological features representation from gray-scale US view; 3) AFF-Net-C, which removes CE gate for enhancement variations modeling.

From Table 2, we observe that fusing deeper-layer features of gray-scale US and contrast-enhanced US provides a clear improvement over the single-path version, with an increase on performance of nearly 7%. Even compared with MC-CNN with early fusion in Table 1, depending on single US modality (A-Net or F-Net) still show inferior performance, further validating the advantage of fusion of morphological features and microvascular perfusion features in our task of thyroid nodule segmentation. When adding channel-wise enhancement gate for explicit perfusion differences representation learning, we could see a significantly higher IoU score (p ¡ 0.05) 69.4% than that of the baseline AFF-Net-C that removes CE gate directly, demonstrating its effectiveness to capture enhancement discrepancy between thyroid nodules and normal gland.

5 Conclusion

In this paper, we have proposed an anatomical-functional fusion network to automatically segment thyroid nodules using dual-screen contrast-enhanced US imaging. Experimental results on our collected datasets have demonstrated the effectiveness of our method in both dynamic enhancement modeling and complementary feature fusion (morphology and perfusion). As the future work, we will extend our current model to a multi-task architecture that jointly detects lesion regions and predicts clinical status for thyroid nodule treatment.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 62136004, 62276130, 61732006, 61876082), and also by the Key Research and Development Plan of Jiangsu Province (No. BE2022842).

References

1. Haugen, B.R., Alexander, E.K., Bible, K.C., et al.: 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid **26**(1), 1–133 (2016)
2. Liang, X.W., Cai, Y.Y., Yu, J.S., Liao, J.Y., Chen, Z.Y.: Update on thyroid ultrasound: a narrative review from diagnostic criteria to artificial intelligence techniques. Chin. Med. J. **132**(16), 1974–1982 (2019)
3. Wang, M., Sun, P., Zhao, X., Sun, Y.: Ultrasound parameters of thyroid nodules and the risk of malignancy: a retrospective analysis. Cancer Control **27**(1), 1073274820945976 (2020)
4. Ha, E.J., Na, D.G., Baek, J.H., Sung, J.Y., Kim, J., et al.: US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. Radiology **287**(3), 893–900 (2018)
5. Kant, R., Davis, A., Verma, V.: Thyroid nodules: advances in evaluation and management. Am. Fam. Physician **102**(5), 298–304 (2020)
6. Sorrenti, S., Dolcetti, V., Fresilli, D., et al.: The role of CEUS in the evaluation of thyroid cancer: from diagnosis to local staging. J. Clin. Med. **10**(19), 4559 (2021)

7. Radzina, M., Ratniece, M., Putrins, D.S., Saule, L., Cantisani, V.: Performance of contrast-enhanced ultrasound in thyroid nodules: review of current state and future perspectives. Cancers **13**(21), 5469 (2021)
8. Ma, J., Wu, F., Jiang, T., et al.: Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. **12**, 1895–1910 (2017)
9. Koundal, D., Sharma, B., Guo, Y.: Intuitionistic based segmentation of thyroid nodules in ultrasound images. Comput. Biol. Med. **121**, 103776 (2020)
10. Mahmood, N.H., Rusli, A.H.: Segmentation and area measurement for thyroid ultrasound image. Int. J. Sci. Eng. Res. **2**(12), 1–8 (2011)
11. Mi, S., Bao, Q., Wei, Z., Xu, F., Yang, W.: MBFF-Net: multi-branch feature fusion network for carotid plaque segmentation in ultrasound. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 313–322. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_30
12. Li, H., et al.: Contrastive rendering for ultrasound image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 563–572. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_54
13. Lu, J., Ouyang, X., Liu, T., Shen, D.: Identifying thyroid nodules in ultrasound images through segmentation-guided discriminative localization. In: Shusharina, N., Heinrich, M.P., Huang, R. (eds.) MICCAI 2020. LNCS, vol. 12587, pp. 135–144. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71827-5_18
14. Lu, J., Yang, J., Batra, D., et al.: Hierarchical question-image co-attention for visual question answering. In: 29th International Proceedings on Advances in Neural Information Processing Systems, Barcelona, Spain. Curran Associates Inc. (2016)
15. Liu, Y., Zhang, X., Zhang, Q., et al.: Dual self-attention with co-attention networks for visual question answering. Pattern Recogn. **117**, 107956 (2021)
16. Aspert, N., Santa-Cruz, D., Ebrahimi, T.: Mesh: measuring errors between surfaces using the hausdorff distance. In: 29th IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, pp. 705–708. IEEE (2002)
17. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., et al.: MMTM: multimodal transfer module for CNN fusion. In: 33th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299. IEEE (2020)
18. Liang, X., Lin, L., Cao, Q., Huang, R., Wang, Y.: Recognizing focal liver lesions in CEUS with dynamically trained latent structured models. IEEE Trans. Med. Imaging **35**(3), 713–27 (2016)
19. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: 32th Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 6087–6096. IEEE (2018)
20. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: 32th Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, pp. 6274–6283. IEEE (2019)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
22. 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

23. Zheng, Q., Delingette, H., Duchateau, N., et al.: 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans. Med. Imaging **37**(9), 2137–2148 (2018)
24. Zhou, S., Wu, H., Gong, J., et al.: Mark-guided segmentation of ultrasonic thyroid nodules using deep learning. In: Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine, pp. 21–26 (2018)
25. Oktay, O., Schlemper, J., Folgoc, L.L., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
26. Lin, G., Shen, C., Van Den Hengel, A., et al.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203 (2016)
27. Chen, L.C., Yang, Y., Wang, J., et al.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
28. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
29. Qin, Y., et al.: Autofocus layer for semantic segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 603–611. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_69
30. Duta, I.C., Liu, L., Zhu, F., et al.: Pyramidal convolution: Rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006.11538 (2020)
31. Ni, J., Wu, J., Tong, J., et al.: GC-Net: global context network for medical image segmentation. Comput. Methods Programs Biomed. **190**, 105121 (2020)
32. Zheng, S., Lu, J., Zhao, H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
33. Kumar, A., Fulham, M., Feng, D., et al.: Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging **39**(1), 204–217 (2019)
34. Zhong, Z., Kim, Y., Zhou, L., et al.: 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: Proceeding of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 228–231. IEEE (2018)
35. Zhao, X., Li, L., Lu, W., et al.: Tumor co-segmentation in PET/CT using multimodality fully convolutional neural network. Phys. Med. Biol. **64**(1), 015011 (2018)
36. Zhang, W., Li, R., Deng, H., et al.: Deep convolutional neural networks for multimodality isointense infant brain image segmentation. Neuroimage **108**, 214–224 (2015)
37. Yang, X., Molchanov, P., Kautz, J.: Multilayer and multimodal fusion of deep neural networks for video classification. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 978–987 (2016)
38. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., et al.: MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299 (2020)

39. Dolz, J., Gopinath, K., Yuan, J., et al.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging **38**(5), 1116–1126 (2018)
40. Li, C., Sun, H., Liu, Z., Wang, M., Zheng, H., Wang, S.: Learning cross-modal deep representations for multi-modal MR image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 57–65. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_7