bioengineering

MDPI

Article

TEM-Net: A Tri-Channel Edge-Aware Multi-Scale Network for Thyroid Nodule Segmentation in Ultrasound Images

Yifei Peng [1,2], Zeru Hai [3,*], Feng Dong [1,2], Bisheng Tang [1,2], Yaoqun Wu [1,2], Xiaoyan Kui [3,*] and Beiji Zou [1,2,3]

[1] School of Information Science and Engineering, Shaoyang University, Shaoyang 422000, China; yfpeng6909@126.com (Y.P.); dongfeng@hnsyu.edu.cn (F.D.); ucas459@gmail.com (B.T.); wuyaoqun@mail.hnsyu.edu.cn (Y.W.); bjzou@csu.edu.cn (B.Z.)
[2] Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang 422000, China
[3] School of Computer Science and Engineering, Central South University, Changsha 410083, China
* Correspondence: haizeru@163.com (Z.H.); xykui@csu.edu.cn (X.K.)

Abstract

With the increasing detection of thyroid nodules in ultrasound screening, accurate nodule segmentation has become important for computer-aided assessment of clinically relevant features, such as contour regularity, aspect ratio, and margin sharpness. Although ultrasound is widely used as a first-line imaging modality in clinical practice, thyroid nodule segmentation remains challenging because of low tissue contrast, speckle-blurred boundaries, and large variations in nodule size and morphology. To address these challenges, we propose TEM-Net, a Tri-Channel Edge-Aware Multi-Scale Network for thyroid nodule segmentation. TEM-Net constructs a tri-channel representation from the raw ultrasound image, including the original grayscale image, a contrast-enhanced image, and a gradient-magnitude map, to highlight weak intensity differences and boundary-related cues in low-contrast ultrasound images. An Edge-Guided Feature Amplification (EGFA) module is introduced before the first down-sampling operation to emphasize boundary responses before spatial resolution is reduced. In addition, a Multi-Focus Cross-Scale Attention Refinement (MF-CAR) module is embedded into skip connections, combining dilated depth-wise convolutions with channel–spatial attention to improve the fusion of local boundary details and broader contextual information. Across three random seeds, TEM-Net achieves mean Dice scores of 0.8822 and 0.9066 and mean IoU scores of 0.7893 and 0.8291 on TN3K and DDTI, respectively, showing competitive performance compared with representative segmentation methods.

Keywords: thyroid nodule segmentation; ultrasound image segmentation; tri-channel contrast–gradient representation; edge-guided feature amplification; cross-scale attention refinement; deep learning

Check for updates

Academic Editor: Fabiano Bini

Received: 27 May 2026
Revised: 6 July 2026
Accepted: 10 July 2026
Published: 15 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Thyroid nodules are frequently detected during ultrasound (US) screening, with previous studies reporting that approximately 50–66% of adults may have at least one thyroid nodule [1]. Although only a small proportion of thyroid nodules are malignant, timely identification of suspicious nodules remains important because advanced thyroid cancer may be associated with extrathyroidal extension or distant metastasis [2]. Ultrasound is widely used as a first-line imaging modality for thyroid nodule assessment because it is non-invasive, cost-effective, widely available, and free of ionizing radiation [3,4]. Clinical

guidelines, including the ACR TI-RADS, incorporate margin characteristics as important criteria for malignancy risk stratification [5,6]. For example, lobulated or irregular margins can increase the TI-RADS score and may influence follow-up or biopsy recommendations. Boundary-related characteristics are relevant to thyroid nodule risk stratification and may influence subsequent clinical assessment [7].

In clinical and research settings, visual assessment and manual delineation of thyroid nodules in grayscale ultrasound images remain operator-dependent and may be subject to inter-observer variability [8,9]. This process is particularly challenging when nodule boundaries are affected by speckle noise, low contrast, or ambiguous surrounding structures [10]. Automated thyroid nodule segmentation may therefore support quantitative measurement, follow-up comparison, and computer-aided biomedical image analysis [11,12].

Deep learning-based segmentation networks, such as U-Net and its variants, have been widely investigated for medical image segmentation and thyroid ultrasound analysis [13–16]. Further developments, such as the DeepLab series, improve receptive-field coverage by using dilated convolutions and atrous spatial pyramid pooling (ASPP) [17]. Transformer-based architectures, such as Swin-UNet [18] and TransUNet [19], have also been explored to model long-range dependencies in segmentation tasks.

Despite these advances, automated thyroid nodule segmentation remains difficult in the presence of several ultrasound-specific image characteristics, as illustrated in Figure 1. First, thyroid nodules often exhibit low contrast against surrounding parenchyma, making intensity-based discrimination difficult for segmentation networks. Second, nodule boundaries can be indistinct or affected by speckle noise, and irregular margins may gradually blend into adjacent tissue, reducing the reliability of edge cues. Third, thyroid nodules vary substantially in size and morphology, ranging from small lesions to large or multifocal nodules with irregular shapes. Such scale and shape diversity makes it difficult for networks with limited receptive-field adaptability to capture fine boundary details and broader contextual information simultaneously.

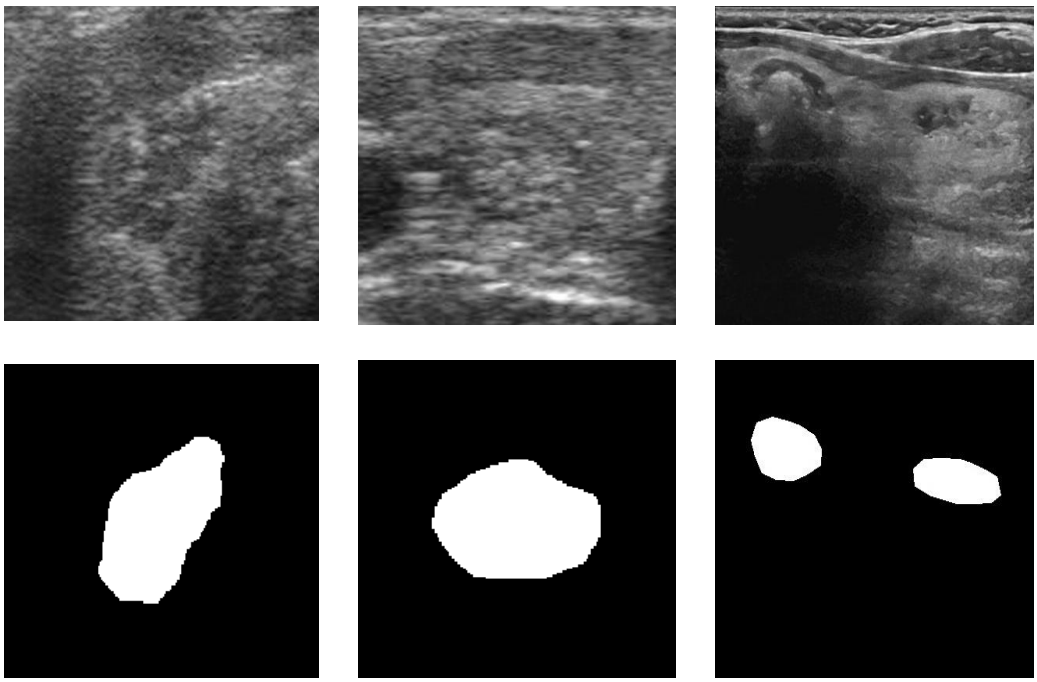


Figure 1. Challenging thyroid ultrasound images and corresponding annotations from the TN3K dataset, illustrating typical difficulties such as low contrast, blurred boundaries, and multiple nodules. In the annotation masks, white regions denote thyroid nodules and black regions denote background.

To address these challenges, we develop TEM-Net and make the following contributions:

(1) Tri-Channel Contrast–Gradient Fusion (TCGF): We construct a three-channel representation from the raw B-mode image, its contrast-enhanced version, and a gradient-based edge map. This representation is designed to expose weak intensity differences and boundary-related cues in low-contrast thyroid ultrasound images.

(2) Edge-Guided Feature Amplification (EGFA): We introduce a Sobel-based gating mechanism before the first down-sampling layer to emphasize boundary responses before spatial resolution is reduced.

(3) Multi-Focus Cross-Scale Attention Refinement (MF-CAR): We embed parallel dilated depth-wise convolutions and channel–spatial attention into skip connections. This design aims to improve the fusion of local boundary details and broader contextual information for nodules with different sizes and shapes.

2. Related Work

Early studies on thyroid nodule analysis in ultrasound images commonly used handcrafted texture features, such as gray-level co-occurrence matrices (GLCM), local binary patterns (LBP), and Gabor filters, together with shallow classifiers or boundary-based models, including support vector machines (SVM), k-nearest neighbors (KNN), and active contours [20–22]. For example, several methods extracted statistical descriptors, such as energy, contrast, and entropy, from GLCM features and used SVM-based models for thyroid nodule classification [23]. However, handcrafted-feature-based models can be sensitive to speckle noise and scanner-dependent gain variations, which may reduce their robustness across different ultrasound devices [24]. Moreover, traditional pipelines usually separate feature extraction from classification or segmentation, limiting end-to-end optimization. This separation may reduce adaptability under conditions of poor boundary definition or low signal intensity [25–27].

Convolutional encoder–decoder networks have promoted the use of deep learning for thyroid nodule segmentation and have improved feature learning compared with handcrafted pipelines [28]. Existing deep learning approaches can be broadly grouped according to their architectural designs. U-Net [29] and its variants, including U-Net++ [30] and Attention U-Net [31], improve encoder–decoder feature fusion through dense skip connections and attention mechanisms. More recently, CE-UNet [32] introduced contrast-sensitive channel attention with sliding-window channel interaction to enhance thyroid nodule segmentation in ultrasound images. Although these models are effective in many medical segmentation tasks, their performance may be affected by thyroid ultrasound-specific factors, including low tissue contrast, speckle-obscured boundaries, and variations in nodule size and shape. To reduce computational cost, lightweight models such as DACNet [33] and DCSAU-Net [34] employ depthwise separable convolutions and compact feature-fusion modules. However, strong compression may limit the ability of some lightweight models to capture fine boundary details and multi-scale contextual information, which are important for delineating nodules with heterogeneous appearances.

Transformer-based architectures have been introduced to model long-range dependencies through self-attention, complementing the local receptive fields of convolutional networks. Hybrid frameworks, including Enhanced-TransUNet [35] and UCTransNet [36], integrate Transformer-based modules into U-Net-like backbones to combine local features with broader contextual representations. Recent methods such as MLMSeg [37] further explore scale-adaptive segmentation by combining convolutional features, Transformer-based contextual modeling, and graph-based structural information. However, self-attention-based designs can introduce considerable computational and memory costs, which may restrict their use in resource-constrained settings. In addition, their performance under

ultrasound-specific artifacts, such as speckle noise and ambiguous margins, still requires further evaluation on thyroid-focused datasets. More recently, state space models (SSMs), such as VM-UNet [38] and UltraLight VM-UNet [39], have been explored as efficient alternatives for medical image segmentation. These architectures are designed to reduce sequence-modeling complexity while maintaining long-range modeling capability. Despite their efficiency, their behavior under common ultrasound imaging challenges, including noise interference, acoustic shadowing, and ambiguous boundaries, remains worth further investigation in complex thyroid nodule cases.

Motivated by these challenges, we propose TEM-Net, a tri-channel edge-aware multi-scale network for thyroid nodule segmentation. TEM-Net constructs a three-channel representation from the raw grayscale image, its contrast-enhanced version, and a gradient-magnitude map to expose intensity, contrast, and boundary-related cues. This contrast–gradient representation is intended to enrich early visual cues for low-contrast regions at the input stage. An edge-guided feature amplification module is applied before the initial down-sampling operation to emphasize boundary responses in regions with weak or indistinct contours. Furthermore, a Multi-Focus Cross-Scale Attention Refinement module is integrated into skip connections to improve the fusion of local boundary details and broader contextual information. By combining contrast–gradient input representation, early edge-guided feature amplification, and cross-scale skip refinement, TEM-Net provides a focused framework for thyroid nodule segmentation in ultrasound images.

3. Method

The overall architecture of TEM-Net is shown in Figure 2. Given an ultrasound frame $I_0 \in \mathbb{R}^{H\times W}$, we first construct a three-channel tensor $X = [I_0, \varepsilon(I_0), g(I_0)]^T \in \mathbb{R}^{3\times H\times W}$, where $\varepsilon(\cdot)$ denotes tile-wise contrast enhancement, and $g(\cdot)$ denotes the normalized gradient magnitude. This TCGF representation provides the network with raw intensity, locally enhanced contrast, and gradient-based edge information while keeping the backbone architecture unchanged.

The tensor X is then processed by an Edge-Guided Feature Amplification (EGFA) module, which produces an edge-aware representation $\widetilde{X} = X + X \odot M$, where $M \in [0,1]^{3\times H\times W}$ is generated from a fixed gradient map through a 1×1 convolution followed by a sigmoid function. EGFA introduces an explicit edge-guided modulation before the first resolution reduction.

The backbone follows a U-Net-like four-stage convolutional encoder–decoder architecture, and its detailed stage-wise configuration is provided in Table 1. Encoder features are passed to the decoder through skip connections, denoted as $\{S_k\}$. Each skip feature S_k is refined by a Multi-Focus Cross-Scale Attention Refinement (MF-CAR) block, formulated as $\widehat{S}_k = S_k + \psi(S_k)$, where $\psi(\cdot)$ aggregates multi-dilated depth-wise responses, applies channel re-scaling, and performs spatial gating with a large-kernel mask. This operation is designed to improve the fusion of local boundary details and broader contextual information, which may help represent small or spatially separated nodules.

The decoder uses bilinear up-sampling and double-convolution blocks to generate the final probability map $\widehat{Y}$. The network is trained with a compound loss, $L = \mathrm{BCE}(\widehat{Y}, Y) + \mathrm{Dice}(\widehat{Y}, Y)$, which combines pixel-wise classification supervision with region-level overlap optimization. Overall, TCGF enriches the input representation with contrast and gradient cues, EGFA emphasizes early boundary responses, and MF-CAR refines skip features across different receptive fields. Together, these components form an edge-aware and cross-scale segmentation framework for thyroid ultrasound images.

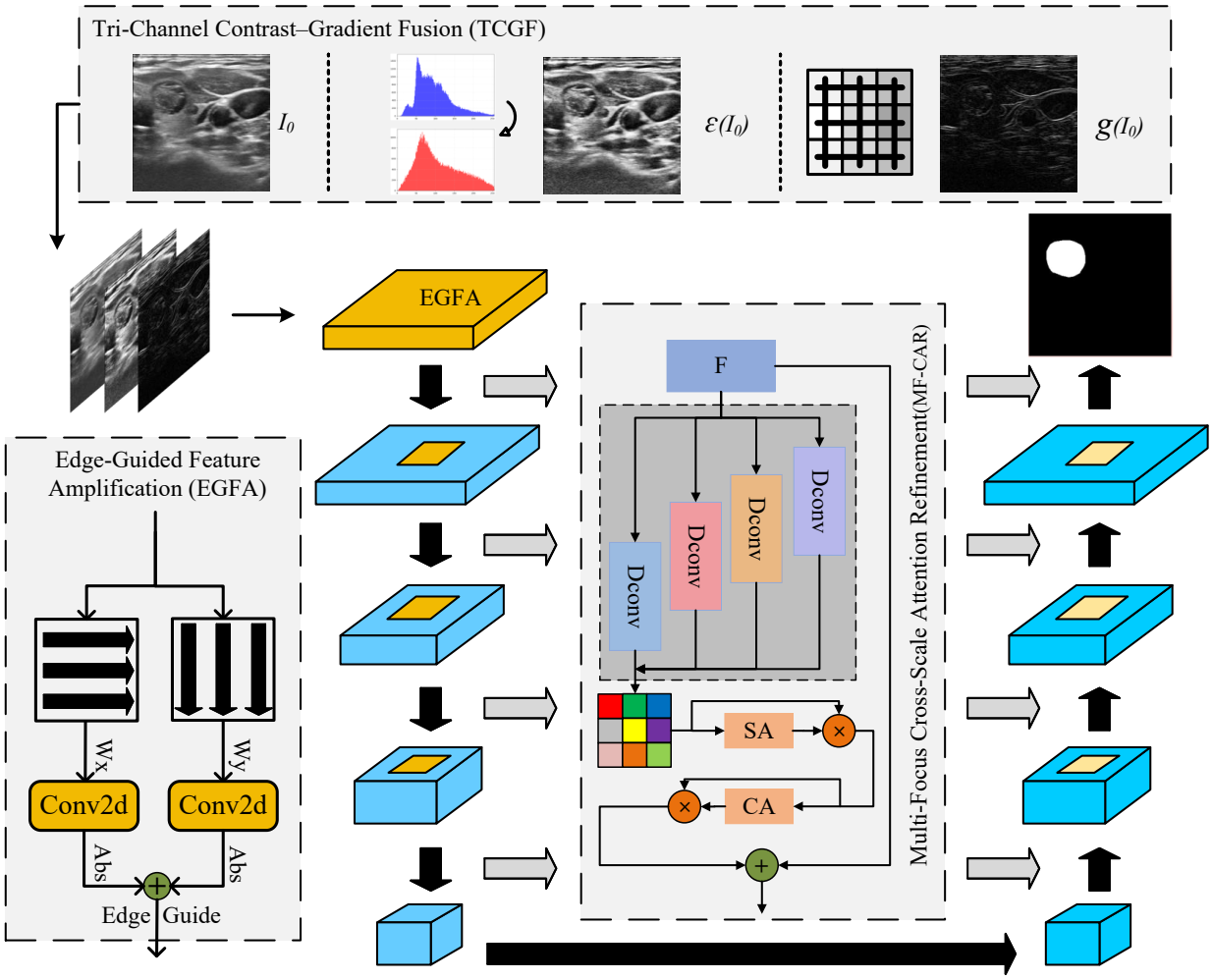


Figure 2. Overall architecture of TEM-Net. The symbol "×" denotes element-wise multiplication, and "+" denotes element-wise addition.

Table 1. Backbone configuration of TEM-Net.

Stage	Resolution	Channels	Main Operation
Input	224 × 224	3	TCGF + EGFA
Stem	224 × 224	64	DC
Encoder 1	112 × 112	128	MP + DC
Encoder 2	56 × 56	256	MP + DC
Encoder 3	28 × 28	512	MP + DC
Bottleneck	14 × 14	512	MP + DC
Decoder 3	28 × 28	256	Up + Concat + DC
Decoder 2	56 × 56	128	Up + Concat + DC
Decoder 1	112 × 112	64	Up + Concat + DC
Decoder 0	224 × 224	64	Up + Concat + DC
Output	224 × 224	1	1 × 1 Conv + Sigmoid

Note: DC denotes two consecutive 3 × 3 convolution–batch normalization–ReLU units. MP denotes 2 × 2 max pooling, and Up denotes bilinear up-sampling. MF-CAR is applied to the deeper encoder features and the bottleneck feature before decoder fusion, while the shallow stem feature is directly connected to the final decoder stage to preserve low-level spatial details.

3.1. Tri-Channel Contrast–Gradient Fusion (TCGF) Input Representation

To improve the representation of low-contrast thyroid ultrasound images, we construct a three-channel tensor from the original B-mode frame and provide it to the encoder as

input. Let $I_0 : \Omega \subset \mathbb{R}^2 \to [0, 255]$ denote the original B-mode frame, where $x \in \Omega$ represents a pixel location. The first channel is the raw ultrasound image, defined as

$$I_{\text{raw}}(x) = I_0(x), \tag{1}$$

which preserves the original echogenic information.

For the second channel, we apply tile-wise contrast-limited adaptive histogram equalization $\varepsilon_{\alpha,T}$, with clip limit $\alpha = 2.0$ and tile size $T = 8$, to locally adjust the gray-level distribution and enhance low-contrast regions:

$$I_{\text{enh}}(x) = \varepsilon_{\alpha,T}[I_0](x). \tag{2}$$

The third channel captures gradient-based boundary information. Spatial derivatives are computed using first-order finite differences, and their Euclidean norm is normalized to the image intensity range:

$$G(x) = \sqrt{(\partial_x I_0)^2 + (\partial_y I_0)^2}. \tag{3}$$

$$I_{\text{grad}}(x) = 255 \cdot \frac{G(x) - \min_\Omega G}{\max_\Omega G - \min_\Omega G + \delta}, \tag{4}$$

where $\delta = 10^{-8}$ is a small constant used for numerical stability.

Unlike the learnable edge-guided module introduced later, I_{grad} is a fixed gradient-derived map that provides explicit boundary-related cues at the input stage. The final three-channel representation is defined as the ordered stack

$$X(x) = \left[I_{\text{raw}}(x),\ I_{\text{enh}}(x),\ I_{\text{grad}}(x)\right]^\top \in \mathbb{R}^3. \tag{5}$$

By combining the raw image, the contrast-enhanced image, and the normalized gradient map, the TCGF representation provides the first convolutional layer with intensity, local-contrast, and boundary-related information without introducing additional trainable parameters.

3.2. Edge-Guided Feature Amplification (EGFA)

The Edge-Guided Feature Amplification (EGFA) module is inserted before the initial down-sampling operation to emphasize boundary-related responses at the early encoding stage, as shown in Figure 3.

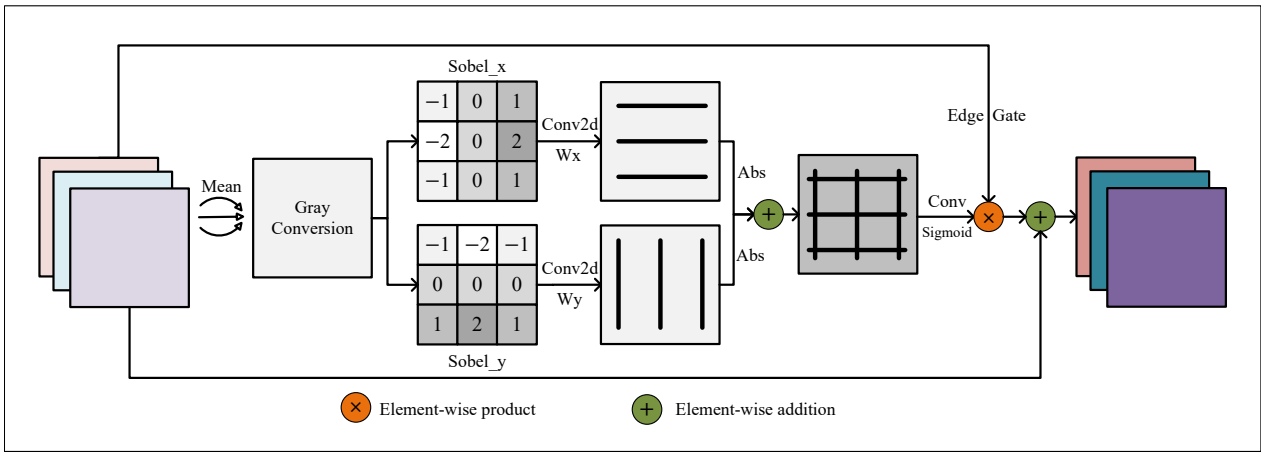


Figure 3. Structure of the Edge-Guided Feature Amplification (EGFA) module.

Given the input tensor $X \in \mathbb{R}^{B\times 3\times H\times W}$, a grayscale map is obtained by channel averaging:

$$I_{\text{gray}} = \frac{1}{3}\sum_{c=1}^{3} X_c. \tag{6}$$

Horizontal and vertical gradients are extracted using fixed Sobel kernels W_x and W_y:

$$W_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad W_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{7}$$

The corresponding gradient responses are computed as

$$G_x = I_{\text{gray}} * W_x, \qquad G_y = I_{\text{gray}} * W_y, \tag{8}$$

where $*$ denotes two-dimensional convolution. An edge-response map is then computed as

$$E = |G_x| + |G_y|. \tag{9}$$

To transform this single-channel edge response into a three-channel modulation mask, E is passed through a 1×1 convolution followed by a sigmoid function:

$$M = \sigma(W_e * E), \qquad W_e \in \mathbb{R}^{3\times 1\times 1\times 1}, \quad M \in (0,1)^{B\times 3\times H\times W}, \tag{10}$$

where W_e denotes a learnable 1×1 convolution kernel that directly projects the single-channel edge response $E \in \mathbb{R}^{B\times 1\times H\times W}$ to a three-channel modulation mask. No intermediate channel is used in this projection.

The output of EGFA is defined as a residual edge-modulated representation:

$$X_{\text{egfa}} = X + X \odot M = X(1 + M), \tag{11}$$

where $\odot$ denotes the Hadamard product. Equation (11) increases features according to the learned edge-response mask, allowing boundary-related regions to receive larger weights while retaining the original residual information.

EGFA introduces only a small number of additional parameters and limited computational overhead because it uses fixed Sobel kernels and a lightweight 1×1 projection. This design incorporates explicit Sobel-based edge information into early feature modulation, which is intended to improve the representation of weak or blurred boundaries in thyroid ultrasound images.

3.3. Multi-Focus Cross-Scale Attention Refinement (MF-CAR) Block

To improve multi-scale feature refinement before decoder fusion, we apply Multi-Focus Cross-Scale Attention Refinement (MF-CAR) blocks to the deeper encoder features and the bottleneck feature. As illustrated in Figure 4, MF-CAR uses parallel dilated depth-wise convolutions followed by branch-specific point-wise projections, feature fusion, and sequential channel–spatial attention to recalibrate feature responses.

Let $F \in \mathbb{R}^{B\times C\times H\times W}$ denote an input feature tensor. Given a dilation set $\mathcal{D} = \{1, 4, 8, 12\}$ with $K = 4$, four branches are applied in parallel. Each branch consists of a 3×3 dilated depth-wise convolution followed by batch normalization and ReLU activation, and then a branch-specific 1×1 point-wise convolution with batch normalization and ReLU activation:

$$F_{d_k} = \rho\Big(\text{BN}\Big(P_{1\times 1}^{(k)}\Big(\rho\Big(\text{BN}\Big(\text{DConv}_{3\times 3}^{(d_k)}(F)\Big)\Big)\Big)\Big)\Big), \quad d_k \in \mathcal{D}, \tag{12}$$

where $\rho(\cdot)$ denotes the ReLU activation, $\mathrm{DConv}_{3\times3}^{(d_k)}$ denotes a 3×3 depth-wise convolution with dilation rate d_k and padding d_k, and $P_{1\times1}^{(k)}$ denotes the branch-specific 1×1 point-wise convolution. In our implementation, the four dilation rates are $d_1 = 1$, $d_2 = 4$, $d_3 = 8$, and $d_4 = 12$.

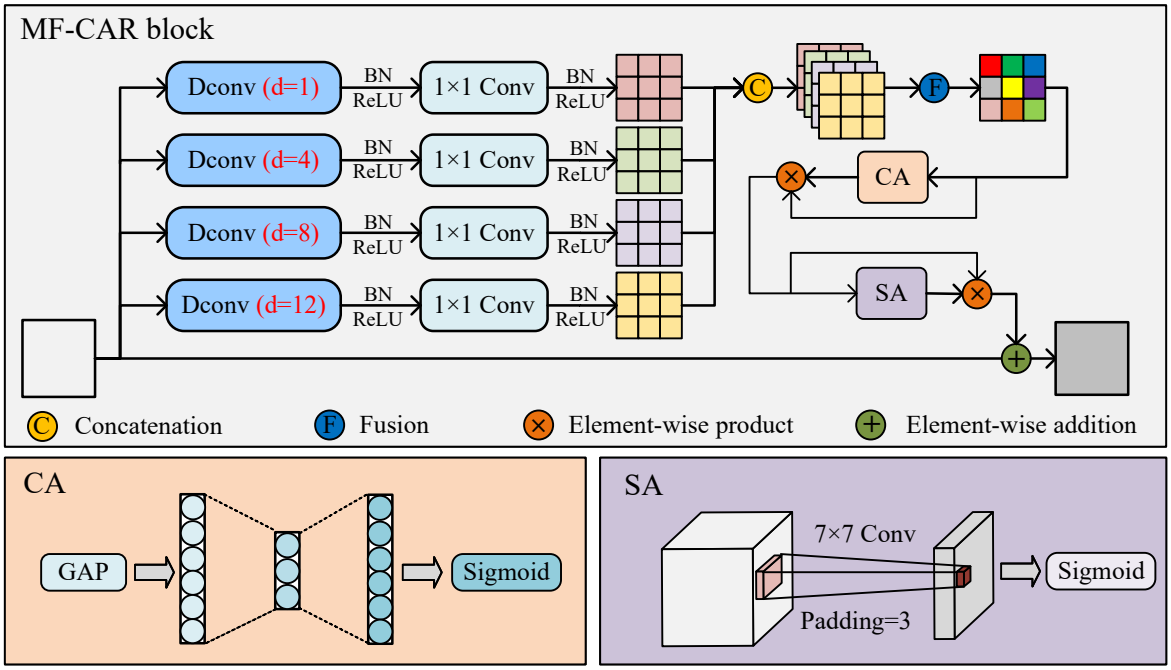


Figure 4. Structure of the Multi-Focus Cross-Scale Attention Refinement (MF-CAR) block.

The branch outputs are concatenated along the channel dimension and further fused by a 1×1 convolution followed by batch normalization and ReLU activation:

$$\widetilde{F} = \rho\Big(\mathrm{BN}\Big(W_f * \big[F_{d_1} \,\|\, \cdots \,\|\, F_{d_K}\big]\Big)\Big), \tag{13}$$

where $\|$ denotes channel-wise concatenation, and $W_f \in \mathbb{R}^{C\times(KC)\times1\times1}$ denotes the 1×1 fusion convolution kernel.

Global average pooling is then applied to produce a channel descriptor $z \in \mathbb{R}^{B\times C}$:

$$z_{b,c} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\widetilde{F}_{b,c,i,j}. \tag{14}$$

The channel attention branch computes channel weights as follows:

$$w_{\mathrm{ch}} = \sigma(W_2\rho(W_1 z)), \quad W_1 \in \mathbb{R}^{\frac{C}{r}\times C}, \quad W_2 \in \mathbb{R}^{C\times\frac{C}{r}}, \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $r = 8$ is the reduction ratio. The channel-refined feature is then obtained as

$$F_{\mathrm{ch}} = \widetilde{F} \odot w_{\mathrm{ch}}, \tag{16}$$

where $\odot$ denotes element-wise multiplication.

A spatial attention map is generated directly from the channel-refined feature using a 7×7 convolution followed by a sigmoid function:

$$w_{\mathrm{sp}} = \sigma(\mathrm{Conv}_{7\times7}(F_{\mathrm{ch}})), \quad w_{\mathrm{sp}} \in (0,1)^{B\times1\times H\times W}. \tag{17}$$

Finally, the spatially reweighted feature is combined with the original input through residual addition:

$$F_{\text{MF-CAR}} = F + F_{\text{ch}} \odot w_{\text{sp}}. \tag{18}$$

3.4. Loss Function

To train the segmentation network with both pixel-wise and region-level supervision, we use a compound loss consisting of binary cross-entropy (BCE) loss and Dice loss. The BCE term provides pixel-wise classification supervision, whereas the Dice term helps account for foreground–background imbalance by directly optimizing region overlap.

Let $p \in [0,1]^N$ denote the flattened predicted probability map, and let $y \in \{0,1\}^N$ denote the corresponding binary ground-truth mask. The total loss is defined as

$$L_{\text{total}} = \lambda_{\text{BCE}} L_{\text{BCE}} + \lambda_{\text{Dice}} L_{\text{Dice}}. \tag{19}$$

where λ_{BCE} and λ_{Dice} are fixed weighting coefficients, both set to 1.0 in our implementation.

The BCE loss is defined as

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]. \tag{20}$$

The Dice loss is expressed as

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \varepsilon}, \tag{21}$$

where $\varepsilon = 10^{-6}$ is used for numerical stability.

This combination jointly considers pixel-wise prediction and region-level overlap, which is useful for thyroid ultrasound images because nodules may occupy a relatively small area and may have weak or blurred boundaries.

4. Experiments

4.1. Datasets

To evaluate the proposed method, we conducted experiments on two publicly available thyroid nodule segmentation datasets: TN3K and DDTI.

The TN3K dataset [40], curated by Gong et al., contains 3493 ultrasound images with pixel-level annotations and was acquired at Zhujiang Hospital of Southern Medical University under ethical approval. Following the commonly used evaluation protocol for TN3K, 614 images were reserved for testing. The remaining 2879 images were used as the training pool, from which 20% (576 images) were randomly selected for validation, leaving 2303 images for training.

The DDTI dataset [41], developed by Pedraza et al., contains 637 ultrasound images with thyroid nodule annotations. For this dataset, we used a random stratified split with a ratio of 7:1:2, resulting in 445 training images, 63 validation images, and 129 testing images.

All images and corresponding segmentation masks were resized to 224×224 pixels to keep the input size consistent during training and evaluation. Comparative experiments were conducted on both TN3K and DDTI, while ablation experiments were performed on TN3K to examine the contribution of each module.

4.2. Implementation Details

All experiments were implemented in PyTorch 1.13 and conducted on a workstation equipped with a single NVIDIA RTX A6000 GPU with 48 GB of memory. TEM-Net was

trained for 300 epochs with a batch size of 32 using the AdamW optimizer [42], with an initial learning rate of 1×10^{-3}. A cosine annealing learning-rate scheduler [43] was used to update the learning rate during training.

Data augmentation was applied during training to reduce overfitting. The augmentation operations included random horizontal flipping, random vertical flipping, and random rotation.

4.3. Evaluation Metrics

To evaluate segmentation performance, we used five commonly adopted metrics: Intersection over Union (IoU), Dice coefficient, accuracy (ACC), specificity (Spe), and sensitivity (Sen). These metrics evaluate region overlap and pixel-wise classification performance from different perspectives. The metrics are computed as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{22}$$

$$\text{Dice} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{23}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{24}$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{25}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{26}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. In addition, we report the number of parameters (Params, M) and the computational complexity measured by floating-point operations (GFLOPs). Statistical significance was assessed using a one-sided paired Wilcoxon signed-rank test based on seed-averaged per-image Dice scores. Specifically, for each test image, the Dice scores obtained from three random seeds were first averaged for each method. The paired Wilcoxon test was then performed between TEM-Net and each compared baseline, with the alternative hypothesis that TEM-Net achieves higher Dice scores. A p-value below 0.05 was considered statistically significant.

4.4. Comparison with Representative Segmentation Methods

We compared TEM-Net with eight representative segmentation methods, including CNN-based, Transformer-based, and Mamba-based models. For a fair comparison, all compared methods were trained for 300 epochs using the same data splits and evaluation protocol. When available, the default configurations from the original implementations were used. To improve reproducibility, the stage-wise backbone configuration of TEM-Net is summarized in Table 1, and the main configuration settings of the compared baseline methods are summarized in Table 2. Quantitative results on TN3K and DDTI are reported in Tables 3 and 4, respectively.

U-Net++ [30] introduces nested and densely connected skip pathways to reduce the semantic gap between encoder and decoder features. DACNet [33] combines depth-wise separable convolutions, squeeze-and-excitation gating, and atrous dual-attention modules to reduce computational cost while preserving multi-scale contextual information. DCSAU-Net [34] uses compact split-attention blocks and asymmetric down-sampling to enhance deep–shallow feature fusion. CE-UNet introduces contrast-sensitive channel attention with sliding-window channel interaction for thyroid nodule segmentation [32].

Table 2. Key configuration settings of the compared baseline models.

Architecture	Algorithm	Configuration Settings
CNN	U-Net++	filters: [32, 64, 128, 256, 512]
CNN	DACNet	filters: [32, 64, 128, 256, 512], split: `fc`, bridge: True
CNN	DCSAU-Net	block: Bottleneck, layers: [2, 2, 2, 2], bottleneck_width: 64, stem_width: 2
CNN	CE-UNet	channels: [64, 128, 256, 512, 1024], CEAM kernels: [3, 5]
Transformer	TransUNet	model: R50-ViT-B_16, activation: `sigmoid`, patch size: 16, token grid: 14 × 14
Transformer	UCTransNet	num_heads: 4, num_layers: 4, patch_sizes: [16, 8, 4, 2], expand_ratio: 4, KV_size: 960
Mamba	Ultra-VM-UNet	c_list: [8, 16, 24, 32, 48, 64], split_att: `fc`, bridge: True
Mamba	VM-UNet	depths: [2, 2, 2, 2], depths_decoder: [2, 2, 2, 1], drop_path_rate: 0.2

Note: All compared models were trained and evaluated using the same 224 × 224 input resolution. For TransUNet, the 14 × 14 token grid is obtained from the 224 × 224 input with a patch size of 16. For UCTransNet, patch sizes [16, 8, 4, 2] are applied to multi-scale encoder features, yielding aligned 14 × 14 token representations across scales.

TransUNet [19] embeds a Vision Transformer encoder into a U-Net-like architecture to model long-range dependencies. UCTransNet [36] replaces conventional skip connections with a Channel Transformer module for multi-scale channel-wise feature fusion. VM-UNet [38] adopts visual state-space blocks to model long-range interactions with reduced sequence-modeling complexity. Ultra-VM-UNet [39] is a compact variant of VM-UNet that introduces parallel Vision-Mamba layers with a small parameter budget.

As shown in Table 3, TEM-Net achieves the highest IoU, Dice, and accuracy among the compared methods on TN3K. Its specificity is comparable to the highest reported value. On DDTI, TEM-Net obtains the highest IoU, Dice, accuracy, and sensitivity. These results indicate strong overlap-based performance under the evaluated in-domain protocols.

Table 3. Quantitative comparison on the TN3K dataset with three random seeds.

Method	Params	FLOPs	IoU	Dice	Acc	Spe	Sen
Ultra-VM-UNet	0.05 M	0.05 G	0.6736 ± 0.0168	0.8049 ± 0.0119	0.9531 ± 0.0036	0.9738 ± 0.0041	0.8019 ± 0.0107
VM-UNet	27.43 M	4.76 G	0.7335 ± 0.0104	0.8462 ± 0.0070	0.9684 ± 0.0074	0.9796 ± 0.0023	0.8424 ± 0.0049
U-Net++	9.16 M	26.51 G	0.7714 ± 0.0064	0.8710 ± 0.0041	0.9694 ± 0.0011	**0.9851 ± 0.0016**	0.8554 ± 0.0082
DACNet	1.28 M	0.56 G	0.7661 ± 0.0050	0.8675 ± 0.0032	0.9681 ± 0.0008	0.9821 ± 0.0013	0.8661 ± 0.0066
DCSAU-Net	2.60 M	5.29 G	0.7781 ± 0.0052	0.8752 ± 0.0033	0.9699 ± 0.0009	0.9833 ± 0.0032	0.8733 ± 0.0200
TransUNet	105.28 M	24.67 G	0.7798 ± 0.0016	0.8763 ± 0.0010	0.9702 ± 0.0007	0.9832 ± 0.0032	0.8753 ± 0.0181
UCTransNet	66.43 M	32.98 G	0.7762 ± 0.0037	0.8740 ± 0.0023	0.9694 ± 0.0007	0.9816 ± 0.0022	**0.8801 ± 0.0117**
CE-UNet	31.04 M	41.91 G	0.7757 ± 0.0063	0.8737 ± 0.0040	0.9697 ± 0.0009	0.9837 ± 0.0011	0.8675 ± 0.0103
TEM-Net	18.54 M	29.52 G	**0.7893 ± 0.0018**	**0.8822 ± 0.0011**	**0.9718 ± 0.0001**	**0.9851 ± 0.0015**	0.8748 ± 0.0105

Note: A one-sided paired Wilcoxon signed-rank test was conducted on seed-averaged per-image Dice scores between TEM-Net and each compared method. All p-values were below 0.05.

Figure 5 and Figure 6 show qualitative comparisons on TN3K and DDTI, respectively. The selected examples include challenging cases with weak boundaries, large nodules, small nodules, multiple nodules, and low contrast. In margin-challenging cases, TEM-Net maintains relatively coherent nodule-region predictions in the selected examples, whereas several compared methods exhibit local irregularities or leakage into adjacent tissue. This observation is consistent with the design motivation of EGFA, which emphasizes early boundary-related responses.

In examples containing small or spatially separated nodules, TEM-Net appears to maintain more coherent coverage of the visible nodule regions than several baseline methods. For large low-contrast lesions, TEM-Net shows fewer fragmented regions and more

coherent overall region predictions. In low-contrast DDTI examples, the contrast-enhanced and gradient-derived channels may help highlight weak boundary-related cues. Overall, the qualitative results are generally consistent with the quantitative trends and suggest that TEM-Net can provide competitive segmentation performance for several challenging ultrasound image characteristics represented in examples.

Table 4. Quantitative comparison on the DDTI dataset with three random seeds.

Method	Params	FLOPs	IoU	Dice	Acc	Spe	Sen
Ultra-VM-UNet	0.05 M	0.05 G	0.7827 ± 0.0129	0.8781 ± 0.0081	0.9296 ± 0.0058	0.9398 ± 0.0104	0.9035 ± 0.0062
VM-UNet	27.43 M	4.76 G	0.8041 ± 0.0144	0.8914 ± 0.0088	0.9381 ± 0.0050	0.9506 ± 0.0045	0.9062 ± 0.0108
U-Net++	9.16 M	26.51 G	0.8136 ± 0.0015	0.8972 ± 0.0009	0.9420 ± 0.0005	0.9566 ± 0.0056	0.9042 ± 0.0131
DACNet	1.28 M	0.56 G	0.8086 ± 0.0065	0.8942 ± 0.0040	0.9400 ± 0.0028	0.9542 ± 0.0055	0.9037 ± 0.0042
DCSAU-Net	2.60 M	5.29 G	0.8198 ± 0.0003	0.9010 ± 0.0002	0.9444 ± 0.0012	$\mathbf{0.9606 \pm 0.0094}$	0.9028 ± 0.0201
TransUNet	105.28 M	24.67 G	0.8185 ± 0.0042	0.9002 ± 0.0026	0.9435 ± 0.0014	0.9570 ± 0.0041	0.9089 ± 0.0101
UCTransNet	66.43 M	32.98 G	0.8213 ± 0.0008	0.9019 ± 0.0005	0.9447 ± 0.0011	0.9574 ± 0.0053	0.9120 ± 0.0097
CE-UNet	31.04 M	41.91 G	0.8179 ± 0.0082	0.8998 ± 0.0049	0.9430 ± 0.0029	0.9543 ± 0.0063	0.9140 ± 0.0131
TEM-Net	18.54 M	29.52 G	$\mathbf{0.8291 \pm 0.0016}$	$\mathbf{0.9066 \pm 0.0010}$	$\mathbf{0.9471 \pm 0.0008}$	0.9597 ± 0.0060	$\mathbf{0.9149 \pm 0.0130}$

Note: A one-sided paired Wilcoxon signed-rank test was conducted on seed-averaged per-image Dice scores between TEM-Net and each compared method. All p-values were below 0.05.

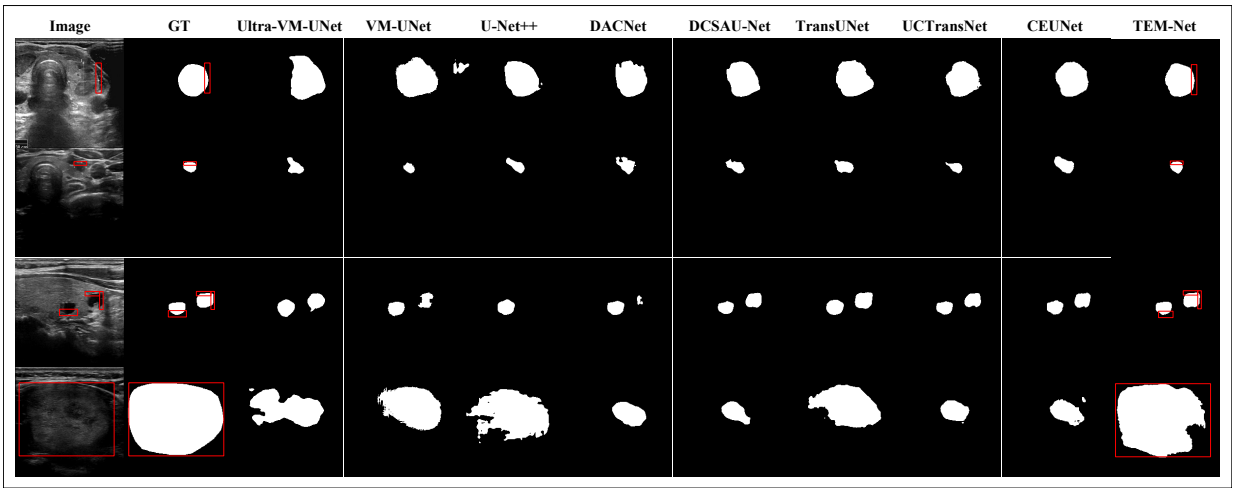


Figure 5. Qualitative comparison of segmentation results on the TN3K dataset. Red boxes indicate challenging regions with visible segmentation differences.

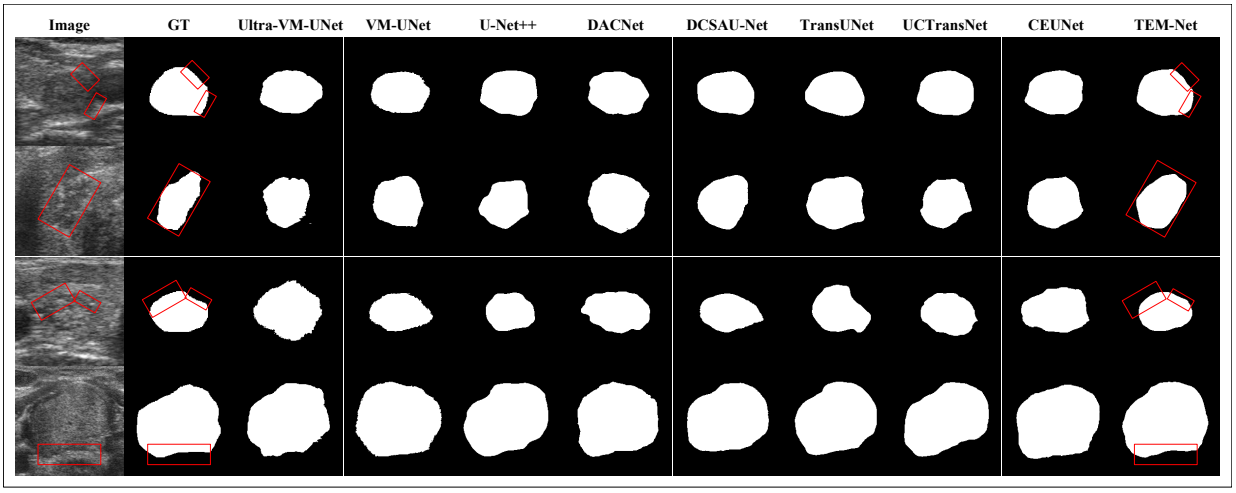


Figure 6. Qualitative comparison of segmentation results on the DDTI dataset. Red boxes indicate challenging regions with visible segmentation differences.

4.5. Ablation Studies

This section presents ablation studies on the TN3K dataset to examine the contribution of each component in TEM-Net. All ablation experiments were conducted under a fixed random seed using the same TN3K split and training protocol.

(1) Effect of TCGF, EGFA, and MF-CAR modules: Table 5 shows the results obtained by progressively adding the proposed components to the baseline model. Compared with the baseline using a single-channel grayscale input, adding TCGF improves IoU from 0.7707 to 0.7770 and Dice from 0.8705 to 0.8745, corresponding to absolute improvements of 0.0063 and 0.0040, respectively. Sensitivity also increases from 0.8536 to 0.8596. These results suggest that the contrast-enhanced image and gradient-derived map provide useful additional cues for low-contrast thyroid ultrasound images.

Table 5. Ablation study of TEM-Net components on the TN3K dataset.

Model	IoU	Dice	Acc	Spe	Sen
Baseline	0.7707	0.8705	0.9693	0.9852	0.8536
Baseline + TCGF	0.7770	0.8745	0.9702	0.9854	0.8596
Baseline + TCGF + EGFA	0.7822	0.8778	0.9709	0.9851	0.8668
Baseline + TCGF + EGFA + MF-CAR	**0.7902**	**0.8828**	**0.9720**	**0.9856**	**0.8732**

Adding EGFA further improves IoU to 0.7822 and Dice to 0.8778. This result is consistent with the design motivation of EGFA, which emphasizes boundary-related responses before the first down-sampling operation. Finally, integrating MF-CAR into skip connections achieves the best IoU and Dice values in this ablation setting, reaching 0.7902 and 0.8828, respectively. Compared with the baseline, the full model improves IoU by 0.0195 and Dice by 0.0123. These results indicate that TCGF, EGFA, and MF-CAR provide cumulative improvements under the TN3K evaluation protocol.

(2) Comparison of MF-CAR with existing attention and multi-scale modules: Table 6 compares MF-CAR with representative attention and multi-scale modules. Compared with CBAM, MF-CAR improves IoU from 0.7843 to 0.7902 and Dice from 0.8791 to 0.8828. Sensitivity also increases from 0.8616 to 0.8732, while specificity decreases slightly from 0.9865 to 0.9856. This comparison indicates that the MF-CAR configuration is associated with higher foreground sensitivity under the evaluated setting.

Table 6. Comparison of MF-CAR with representative attention and multi-scale modules.

Module	Params	FLOPs	Epoch Time	IoU	Dice	Acc	Spe	Sen
CBAM	13.70 M	23.83 G	27.67 s	0.7843	0.8791	0.9714	0.9865	0.8616
Lite-ASPP	17.06 M	27.23 G	60.50 s	0.7879	0.8813	0.9718	0.9859	0.8690
ASPP	38.87 M	51.27 G	72.83 s	**0.7906**	**0.8831**	**0.9722**	0.9862	0.8699
MF-CAR	18.54 M	29.52 G	40.83 s	0.7902	0.8828	0.9720	0.9856	**0.8732**

Compared with Lite-ASPP, MF-CAR achieves slightly higher IoU (0.7902 vs. 0.7879) and Dice (0.8828 vs. 0.8813), with moderate increases in parameters and FLOPs. Compared with the full ASPP module, however, MF-CAR is marginally lower in IoU (0.7902 vs. 0.7906), Dice (0.8828 vs. 0.8831), accuracy (0.9720 vs. 0.9722), and specificity (0.9856 vs. 0.9862), while achieving slightly higher sensitivity (0.8732 vs. 0.8699). Thus, ASPP remains marginally better when overlap accuracy alone is prioritized.

Nevertheless, MF-CAR reduces the parameter count from 38.87 M to 18.54 M and FLOPs from 51.27 G to 29.52 G, corresponding to reductions of approximately 52.3% and 42.4%, respectively. Under the same experimental setting, the average epoch-level wall-clock time was 40.83 s for MF-CAR and 72.83 s for ASPP. Overall, these results indicate

that MF-CAR provides a performance–efficiency trade-off rather than a uniform accuracy advantage over ASPP.

(3) Effect of hybrid loss function: Compared with BCE loss alone, adding Dice loss slightly improves the overlap-based metrics. As shown in Table 7, IoU increases from 0.7851 to 0.7902, Dice increases from 0.8796 to 0.8828, and sensitivity increases from 0.8676 to 0.8732. Specificity remains unchanged at 0.9856, while accuracy increases slightly from 0.9713 to 0.9720.

Table 7. Comparison between BCE loss and the hybrid BCE + Dice loss.

Loss	IoU	Dice	Acc	Spe	Sen
BCE	0.7851	0.8796	0.9713	**0.9856**	0.8676
BCE + Dice	**0.7902**	**0.8828**	**0.9720**	**0.9856**	**0.8732**

Bold values indicate the best performance among the compared loss settings.

These results suggest that the Dice term helps account for foreground–background imbalance by directly optimizing region overlap. The improvement is modest but consistent across IoU, Dice, and sensitivity in this experiment. Therefore, the BCE+Dice loss is used as the default training objective in TEM-Net.

5. Discussion

On the two public thyroid nodule segmentation datasets used in this study, TN3K and DDTI, TEM-Net achieves the highest Dice and IoU scores among the compared methods. It uses 18.5 M parameters and 29.52 GFLOPs, indicating a moderate parameter count but a non-negligible computational cost compared with lightweight alternatives. Qualitative results further suggest that TEM-Net can produce relatively coherent nodule-region predictions in several challenging examples, such as weak-boundary nodules, low-contrast lesions, and small or multiple nodules.

These results are consistent with the design motivation of the three proposed modules. TCGF provides the first convolutional layer with raw-intensity, local-contrast, and gradient-derived boundary cues. EGFA emphasizes edge-related responses before the first resolution reduction. MF-CAR refines skip features through dilated depth-wise convolutions and channel–spatial attention, which may improve the fusion of local boundary details and broader contextual information. Together, these components provide an edge-aware and cross-scale framework for thyroid ultrasound image segmentation.

It should be noted that Dice and IoU quantify regional overlap between model predictions and reference annotations rather than direct clinical utility. This distinction is particularly relevant in thyroid ultrasound, where manual nodule delineation is observer-dependent [8,9]. Across three random seeds, TEM-Net achieved Dice/IoU values of $0.8822 \pm 0.0011/0.7893 \pm 0.0018$ on TN3K and $0.9066 \pm 0.0010/0.8291 \pm 0.0016$ on DDTI. The small variations across runs indicate stable overlap with the available reference annotations. These results support the potential use of TEM-Net as an automated segmentation component for region-level quantitative analysis, while multi-reader studies and direct comparison with inter-observer variability are still needed to establish clinical utility.

Several limitations warrant further investigation. First, the current pipeline processes only single B-mode ultrasound images. Incorporating additional ultrasound information, such as Doppler flow or elastography, may be useful for future studies involving more comprehensive thyroid nodule assessment [44]. Second, the local-contrast operator in TCGF uses fixed hyperparameters. Future work could investigate learnable or content-adaptive enhancement strategies to improve adaptability across imaging conditions. Third, cross-dataset transferability remains a clear limitation of the current study. The strong in-

domain performance observed on TN3K and DDTI should not be interpreted as robustness to datasets with entirely unseen distributions. Future work should investigate dedicated domain adaptation and domain generalization strategies to further improve cross-dataset transferability. Fourth, although TEM-Net has a moderate number of parameters, its FLOP cost is higher than that of several lightweight models, and further model compression or inference-time analysis would be required for deployment-oriented applications. Finally, uncertainty estimation may help identify low-confidence predictions and support more cautious interpretation in ambiguous cases [45].

6. Conclusions

We introduced TEM-Net, a tri-channel edge-aware multi-scale network for thyroid nodule segmentation in ultrasound images. The method combines contrast–gradient input representation, early edge-guided feature amplification, and cross-scale skip-feature refinement. Experiments under the evaluated in-domain protocols on TN3K and DDTI show that TEM-Net achieves competitive Dice and IoU scores among the compared methods. These findings support TEM-Net as a competitive segmentation framework under the evaluated in-domain protocols, while broader cross-dataset generalizability remains to be established.

Author Contributions: Conceptualization, Y.P., Z.H. and X.K.; methodology, Y.P. and Z.H.; validation, F.D.; investigation, Y.P. and F.D.; resources, Y.P. and B.Z.; data curation, F.D., B.T. and Y.W.; writing—original draft preparation, Y.P. and Z.H.; writing—review and editing, Y.P., Z.H., B.T., Y.W., X.K. and B.Z.; supervision, Z.H., X.K. and B.Z.; funding acquisition, Y.P., F.D., X.K. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number U22A2034; the Scientific Research Foundation of Hunan Provincial Education Department, grant numbers 24A0018 and 22A0530; the Natural Science Foundation of Hunan Province, grant numbers 2024JJ7508 and 2024JJ7496; the Major Program from Xiangjiang Laboratory, grant number 23XJ02005; the Central South University Research Programme of Advanced Interdisciplinary Studies, grant number 2023QYJC020; and the Shaoyang Key Research and Development Innovation Project, grant number 2025GZ2025.

Institutional Review Board Statement: Ethical review and approval were waived for this study because only publicly available, anonymized datasets were used, and no new human participant data were collected.

Data Availability Statement: The data analyzed in this study are publicly available thyroid ultrasound datasets. The TN3K and DDTI datasets can be obtained from their original dataset providers or corresponding publications. No new original dataset was generated in this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of this manuscript; or in the decision to publish the results.

References

1. Kobaly, K.; Kim, C.S.; Mandel, S.J. Contemporary management of thyroid nodules. *Annu. Rev. Med.* **2022**, *73*, 517–528. [CrossRef] [PubMed]
2. Nguyen, Q.T.; Lee, E.J.; Huang, M.G.; Park, Y.I.; Khullar, A.; Plodkowski, R.A. Diagnosis and treatment of patients with thyroid cancer. *Am. Health Drug Benefits* **2015**, *8*, 30. [PubMed]
3. Fresilli, D.; David, E.; Pacini, P.; Del Gaudio, G.; Dolcetti, V.; Lucarelli, G.T.; Di Leo, N.; Bellini, M.I.; Pacini, F.; Cantisani, V. Thyroid nodule characterization: How to assess the malignancy risk. Update of the literature. *Diagnostics* **2021**, *11*, 1374. [CrossRef] [PubMed]
4. Hoang, J.K.; Oldan, J.D.; Mandel, S.J.; Orloff, L.A.; Bauman, G.S.; Berman, C.G.; Danya, A.J.; Gurgel, R.K.; Ho, A.S.; Lee, S.L.; et al. ACR appropriateness criteria® thyroid disease. *J. Am. Coll. Radiol.* **2019**, *16*, S300–S314. [CrossRef] [PubMed]

5. Hoang, J.K.; Middleton, W.D.; Tessler, F.N. Update on ACR TI-RADS: Successes, challenges, and future directions. *AJR Amer. J. Roentgenol.* **2021**, *216*, 570–578. [CrossRef]
6. Tappouni, R.R.; Itri, J.N.; McQueen, T.S.; Lalwani, N.; Ou, J.J.; Lee, E.Y. ACR TI-RADS: Pitfalls, solutions, and future directions. *Radiographics* **2019**, *39*, 2040–2052. [CrossRef] [PubMed]
7. Tessler, F.N.; Middleton, W.D.; Grant, E.G. Thyroid imaging reporting and data system (TI-RADS): A user's guide. *Radiology* **2018**, *287*, 29–36. [CrossRef] [PubMed]
8. Choi, S.H.; Kim, E.K.; Kwak, J.Y.; Kim, M.J.; Son, E.J.; Oh, K.K.; Kim, S.H.; Lee, S.H. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* **2010**, *20*, 167–172. [CrossRef] [PubMed]
9. Alyami, J.; Almutairi, F.F.; Aldoassary, S.; Alharbi, A.F.; Alharthi, S.M.; Alqarni, A.S.; Alharthi, A.M.; Alharbi, M.A.; Alqahtani, A.A.; Alshehri, A.M. Interobserver variability in ultrasound assessment of thyroid nodules. *Medicine* **2022**, *101*, e31106. [CrossRef] [PubMed]
10. Gong, H.; Chen, G.; Wang, R.; Xie, X.; Mao, M.; Yu, Y.; Chen, F.; Li, G. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021*; IEEE: New York, NY, USA, 2021; pp. 257–261.
11. Kang, Q.; Lao, Q.; Li, Y.; Li, J.; Xu, Y.; Liu, Y.; Zhang, Y.; Sun, L. Thyroid nodule segmentation and classification in ultrasound images through intra-and inter-task consistent learning. *Med. Image Anal.* **2022**, *79*, 102443. [CrossRef] [PubMed]
12. Hu, M.; Zhang, Y.; Xue, H.; Li, Y.; Wang, J.; Chen, L.; Liu, S. Mamba-and resnet-based dual-branch network for ultrasound thyroid nodule segmentation. *Bioengineering* **2024**, *11*, 1047. [CrossRef] [PubMed]
13. Ibtehaz, N.; Kihara, D. Acc-UNet: A completely convolutional UNet model for the 2020s. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer Nature: Cham, Switzerland, 2023; pp. 692–702.
14. Zhou, H.; Luo, Y.; Guo, J.; Chen, Z.; Gong, W.; Lin, Z.; Zhuo, M.; Lin, Y.; Lin, W.; Shen, Q. Double U-Net: Semi-supervised ultrasound image segmentation combining CNN and transformer's U-shaped network. *J. Supercomput.* **2025**, *81*, 659. [CrossRef]
15. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. UNet 3+: A full-scale connected UNet for medical image segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2020; pp. 1055–1059.
16. Ji, Z.; Zou, B.; Kui, X.; Li, H.; Vera, P.; Ruan, S. Mamba-based feature extraction and adaptive multilevel feature fusion for 3D tumor segmentation from multi-modal medical image. *arXiv* **2025**, arXiv:2504.21281.
17. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
18. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, J.; Tian, Q.; Wang, M. Swin-UNet: UNet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer Nature: Cham, Switzerland, 2022; pp. 205–218.
19. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
20. Guetari, R.; Ayari, H.; Sakly, H. Computer-aided diagnosis systems: A comparative study of classical machine learning versus deep learning-based approaches. *Knowl. Inf. Syst.* **2023**, *65*, 3881–3921. [CrossRef]
21. Vadhiraj, V.V.; Simpkin, A.; O'Connell, J.; O'Mahony, N.; Murphy, S.; O'Connor, M.; O'Sullivan, J.N. Ultrasound image classification of thyroid nodules using machine learning techniques. *Medicina* **2021**, *57*, 527. [CrossRef] [PubMed]
22. Acharya, U.R.; Chowriappa, P.; Fujita, H.; Bhat, S.; Dua, S.; Koh, J.E.W.; Tan, J.H.; Chua, K.C.; Ng, K.H. Thyroid lesion classification using Gabor transform features from high resolution ultrasound images. *Knowl.-Based Syst.* **2016**, *107*, 235–245. [CrossRef]
23. Dong, P.; Zhang, R.; Li, J.; Wang, L.; Chen, Y.; Liu, X.; Zhao, H. An ultrasound image segmentation method for thyroid nodules based on dual-path attention mechanism-enhanced UNet++. *BMC Med. Imag.* **2024**, *24*, 341. [CrossRef]
24. Singh, P. Feature Enhanced Speckle Reduction in Ultrasound Images: Algorithms for Scan Modelling, Speckle Filtering, Texture Analysis and Feature Improvement. Ph.D. Thesis, University of Canterbury, Christchurch, New Zealand, 2019.
25. Abdolali, F.; Shahroudnejad, A.; Amiri, S.; Hareendranathan, A.R.; Jaremko, J.L.; Noga, M.; Punithakumar, K. A systematic review on the role of artificial intelligence in sonographic diagnosis of thyroid cancer: Past, present and future. *Front. Biomed. Technol.* **2020**, *7*, 266–280.
26. Garg, M.; Dhiman, G. A novel content-based image retrieval approach using GLCM features and texture fused LBP variants. *Neural Comput. Appl.* **2021**, *33*, 1311–1328.
27. Bi, Y.; Jiang, Z.; Duelmer, F.; Huang, D.; Navab, N. Machine learning in robotic ultrasound imaging: Challenges and perspectives. *Annu. Rev. Control Robot. Auton. Syst.* **2024**, *7*, 335–357. [CrossRef]
28. Etehadtavakol, M.; Etehadtavakol, M.; Ng, E.Y.K. Enhanced thyroid nodule segmentation through U-Net and VGG16 fusion with feature engineering: A comprehensive study. *Comput. Methods Programs Biomed.* **2024**, *251*, 108209. [CrossRef] [PubMed]

29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
30. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J.; Chen, Y.; Wu, D.; Wang, C.; Li, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Proceedings of the DLMIA and ML-CDS Workshops, MICCAI*; Springer: Granada, Spain, 2018; pp. 3–11.
31. Oktay, O.; Schlemper, J.; Le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
32. Zhang, M.; Ma, M.; Zhao, B.; Li, H.; Wang, X.; Chen, Y.; Liu, J. CE-UNet: A contrast-sensitive channel attention network with sliding-window interaction for thyroid nodule segmentation. *Biomed. Signal Process. Control* **2026**, *119*, 109880. [CrossRef]
33. Yang, Y.; Huang, H.; Shao, Y.; Li, J.; Wang, X.; Zhang, L.; Chen, Q. DAC-Net: A light-weight U-shaped network based efficient convolution and attention for thyroid nodule segmentation. *Comput. Biol. Med.* **2024**, *180*, 108972. [CrossRef] [PubMed]
34. Xu, Q.; Ma, Z.; Duan, W. DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. *Comput. Biol. Med.* **2023**, *154*, 106626. [CrossRef] [PubMed]
35. Ozcan, A.; Tosun, Ö.; Donmez, E.; Yilmaz, B.; Kaya, M.; Demir, S. Enhanced-TransUNet for ultrasound segmentation of thyroid nodules. *Biomed. Signal Process. Control* **2024**, *95*, 106472. [CrossRef]
36. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.; Liu, J.; Wang, Y.; Li, J. UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2441–2449. [CrossRef]
37. Chen, G.; Tan, G.; Duan, M.; Li, Y.; Wang, H.; Zhang, X.; Liu, S. MLMSeg: A multi-view learning model for ultrasound thyroid nodule segmentation. *Comput. Biol. Med.* **2024**, *169*, 107898. [CrossRef] [PubMed]
38. Ruan, J.; Li, J.; Xiang, S. VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv* **2024**, arXiv:2402.02491.
39. Wu, R.; Liu, Y.; Liang, P.; Zhang, J.; Li, Q.; Wang, Z.; Chen, X. UltraLight VM-UNet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv* **2024**, arXiv:2403.20035.
40. Gong, H.; Chen, J.; Chen, G.; Li, H.; Li, G.; Chen, F. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput. Biol. Med.* **2023**, *155*, 106389. [CrossRef] [PubMed]
41. Pedraza, L.; Vargas, C.; Narváez, F.; Durán, O.; Muñoz, E.; Romero, E. An open access thyroid ultrasound image database. In *Proceedings of the 10th International Symposium on Medical Information Processing and Analysis*; SPIE: Bellingham, WA, USA, 2015; pp. 188–193.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
43. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
44. Ansari, M.Y.; Mangalote, I.A.C.; Meher, P.K.; Aboumarzouk, O.; Al-Ansari, A.; Halabi, O.; Dakua, S.P. Advancements in deep learning for B-mode ultrasound segmentation: A comprehensive review. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2126–2149. [CrossRef]
45. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5574–5584.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.