

Attention Augmented U-Net for Robust Thyroid Ultrasound Segmentation

Özgür Ömer Gülsoy¹

ozguromergulsoy@gmail.com

Istanbul Okan University,
Computer Engineering, Istanbul,
Turkey¹

Mohammed Qusay Mawlood²

mohammawlood@stu.okan.edu.tr

Istanbul Okan University,
Artificial Engineering, Istanbul,
Turkey²

Mahsa Mikaeili³

mahsa.mikaeili@okan.edu.tr

Istanbul Okan University,
Mechatronic Engineering,
Istanbul, Turkey³

Sina Alp⁴

Sina.alp@okan.edu.tr

Istanbul Okan University, Electric
and Electronics Engineering,
Istanbul, Turkey⁴

Abstract—Thyroid nodule segmentation in ultrasound imaging faces challenges from speckle noise, low contrast, and morphological variability. We propose an EfficientNet-B5-backed UNet++ architecture enhanced with Dual-Path Attention Modules (DPAM) integrating Global Context, Efficient Channel, and Spatial Attention mechanisms. Our model employs a composite loss function combining Focal Tversky, Dice, and Boundary losses with temporal ramp-up strategy to address class imbalance. Training features a three-phase progressive schedule with automatic threshold optimization and progressive resizing to 512×512, complemented by ultrasound-specific data augmentations. Evaluated on the TN3K dataset using 5-fold cross-validation, our model achieves state-of-the-art performance: 0.8875 Dice, 0.7969 IoU, 0.9688 accuracy, representing improvement over recent CNN and Transformer baselines while maintaining computational efficiency with less than 50M parameters. The tight training-validation gaps demonstrate robust generalization, indicating reliable boundary delineation and clinical applicability for computer-aided diagnosis pipelines.

Keywords—Thyroid Segmentation, Ultrasound Imaging, UNet++, Dual-Path Attention Module, Medical Image Analysis

I. INTRODUCTION (HEADING 1)

Thyroid nodules rank among the most common endocrine disorders, with B-mode ultrasound (US) serving as the primary modality for detection, characterization, and follow-up due to

its real-time capability, low cost, and safety. However, US images suffer from speckle noise, low contrast, heterogeneous texture, and ill-defined lesion borders, making manual delineation time-consuming and operator-dependent. Accurate automated segmentation is therefore critical for clinical workflows and computer-aided diagnosis (CAD) systems [1,2,3].

UNet++ with redesigned skip connections for multiscale feature fusion [5], ResUNet with residual encoders [1], and attention-augmented variants that dynamically recalibrate feature maps [6,7]. EfficientNet backbones [8] have proven effective in data-scarce scenarios, while Transformer-based hybrids like Enhanced-TransUNet [7] and BPAT- feature extraction. Addressing thyroid ultrasound's unique challenges—blurred boundaries and severe class imbalance—recent methods employ boundary-aware supervision, dual-path attention, shape priors, and composite loss functions. Focal Tversky [13], boundary loss [11], and integrated multi-component losses like those in DPAM-UNet++ [10] have demonstrated superior edge precision compared to standard Dice and cross-entropy losses.

This work advances the field by combining an EfficientNet-B5-backed UNet++ architecture with Dual-Path Attention Mechanism (DPAM), trained using a composite loss integrating Focal Tversky, Dice, and boundary terms. Our pipeline

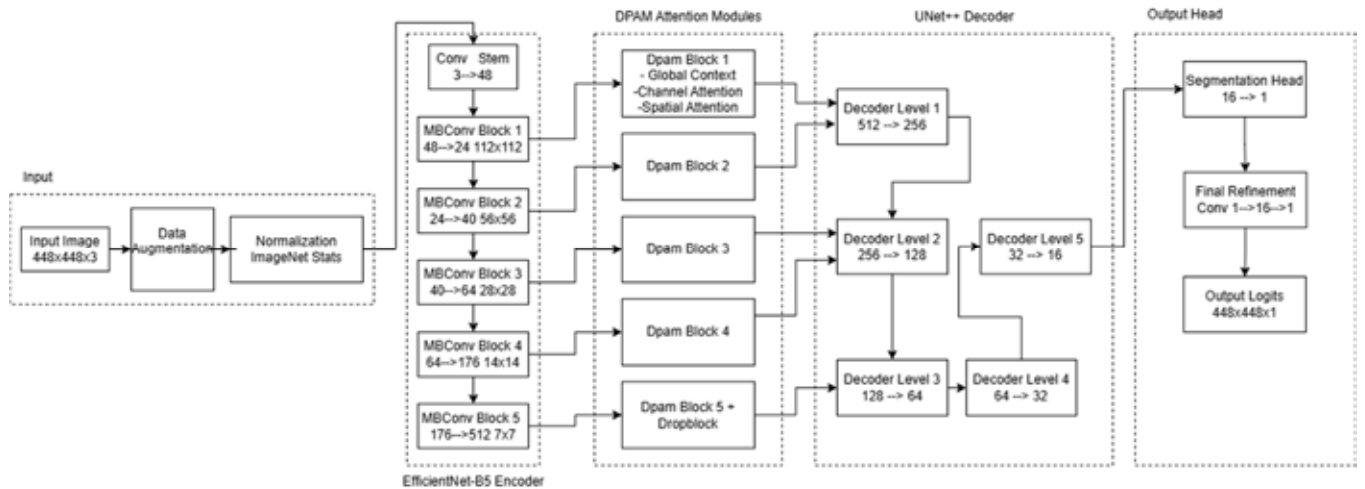


Fig. 1. Proposed models architecture diagram.

incorporates domain-specific augmentations and a three-stage training schedule: frozen encoder initialization, full fine-tuning, and progressive resizing. The paper is organized as follows: Section II presents the proposed methodology, Section III discusses experimental results, and Section IV provides conclusions.

II. METHODS AND MATERIALS

A. Model Architecture

Our model employs EfficientNet-B5 as an encoder within a UNet++ architecture, chosen for its strong feature extraction capabilities and excellent segmentation performance in medical imaging tasks. The key enhancement of our architecture as shown in Fig. 1, is the strategic integration of a Dual-Path Attention Module (DPAM) after each encoder stage except the initial stem stage resulting in five DPAM blocks at critical feature extraction points. Each DPAM combines three complementary attention mechanisms: Global Context Attention [15] for capturing long-range dependencies, Efficient Channel Attention [16] for modeling inter-channel relationships with minimal computational overhead, and Spatial Attention [17] for emphasizing diagnostically relevant regions while suppressing background noise. These attention pathways operate in parallel, allowing simultaneous focus on different ultrasound feature aspects. The outputs from the three attention mechanisms are concatenated and merged using 1×1 convolutional layers with batch normalization, effectively learning optimal feature combinations. The decoder leverages UNet++'s densely connected skip connections with nested pathways, which help maintain fine-grained spatial details and ensure efficient gradient flow—both essential for accurate boundary detection in ultrasound images. Finally, a refinement module with additional convolutional layers further processes the segmentation logits, producing precise binary segmentation masks that match the original 448×448 pixel input size.

B. Loss Function

In this study, we employ an Enhanced Composite Loss that integrates multiple specialized losses: Focal Tversky Loss [13], Dice Loss, and Boundary Loss [11]. The composite loss function is defined as:

$$L = \alpha_c L_{FT} + \beta_c L_{Dice} + \gamma_c(t) L_{Boundary} \quad (1)$$

Where L_{FT} is the Focal Tversky Loss, L_{Dice} is the Dice Loss, and $L_{Boundary}$ is the Boundary Loss. The coefficients α_c , β_c , and $\gamma_c(t)$ balance these components. Focal Tversky Loss (L_{FT}): This loss emphasizes challenging pixels, formulated as:

$$L_{FT} = \left[1 - \frac{TP}{TP + \alpha FN + \beta FP + \epsilon} \right]^\gamma \quad (2)$$

Where: TP is the true positive, FN is the false negative, FP is the false positive parameter, α and β control sensitivity to false negatives and false positives, respectively. We set $\alpha=0.7$, $\beta=0.3$, and $\gamma=1.5$ to enhance attention on difficult-to-segment regions. ϵ is a small constant ($1e^{-6}$) ensuring numerical stability. Dice Loss (L_{Dice}): Dice Loss ensures good overlap between predicted masks and ground truth masks, defined as:

$$L_{Dice} = 1 - \frac{2 \times (P \cap T)}{|P| + |T| + \epsilon} \quad (3)$$

where P is the predicted segmentation and T is the ground truth. Boundary Loss ($L_{Boundary}$): This loss focuses explicitly on the boundaries between segmented regions, calculated using gradient edge maps derived using Sobel filters:

$$L_{Boundary} = \frac{1}{N} \sum_{i=1}^N |\nabla P_i - \nabla T_i| \quad (4)$$

where ∇P and ∇T are gradient maps of predictions and targets, respectively, and N denotes the total number of pixels. To balance region segmentation and boundary precision dynamically during training, we employ a ramp-up strategy for the boundary loss coefficient $\gamma_c(t)$:

$$\gamma_c(t) = \gamma_{start} + (\gamma_{end} - \gamma_{start}) \times \min\left(1, \frac{t}{t_{ramp}}\right) \quad (5)$$

Where: γ_{start} and γ_{end} define initial and maximum weights of boundary loss. In this study, these are set as 0.15 and 0.35 respectively. T is the current training epoch, and t_{ramp} is the epoch threshold to reach the maximum boundary loss contribution set as 20 epochs.

C. Datasets

In our experiments, we utilize the open-access Thyroid Nodule TN3K dataset [14], which is specifically curated for thyroid nodule segmentation tasks. The TN3K dataset consists of 3,493 B-mode ultrasound images with corresponding pixel-wise nodule annotations, collected from 2,421 patients at Zhujiang Hospital of Southern Medical University. The images were acquired between January 2016 and August 2020 using three different ultrasound systems—GE Logiq E9, Hitachi ARIETTA 850, and Mindray RESONA 70B—to ensure diversity in imaging equipment and acquisition settings.

III. EXPERIMENTS

A. Training and Evaluation

To evaluate model generalization, we conduct five-fold cross-validation. The dataset is divided into five folds, each containing approximately 698–699 images. In each iteration, four folds (2,794 images) are used for training, while the remaining fold (699 images) is used for validation. We use metrics such as Dice, IoU, Accuracy, Precision and Recall for our performance metrics. Across all five folds, our model achieves the following mean on the held-out validation sets Fig.2 and Table.1. These results demonstrate consistency across folds, with sub-1% standard deviations on all primary metrics.

Fig. 2 shows the evolution of Dice, IoU and accuracy on both the training and validation sets over 48 epochs. We employ a three-phase training strategy with 5-fold cross-validation. Phase 1: 5 epochs with frozen encoder, AdamW ($lr=1e-4$, $wd=1e-4$), 5-epoch warmup followed by cosine annealing. Phase 2: 35 epochs with all layers unfrozen, automatic threshold search every 5 epochs starting at epoch 15, early stopping (patience=10). Phase 3: 8 epochs at 512×512 resolution ($lr=1e-5$, batch size 8), threshold search every 2 epochs. All phases use morphological post-processing (opening kernel 3, closing kernel 5) and retain the largest

Table.1. Performance on validation data.

Metric	Mean
Dice	0.8875
IoU	0.7969
Accuracy	0.9688
Precision	0.8827
Recall	0.8955
Optimal Threshold	0.544

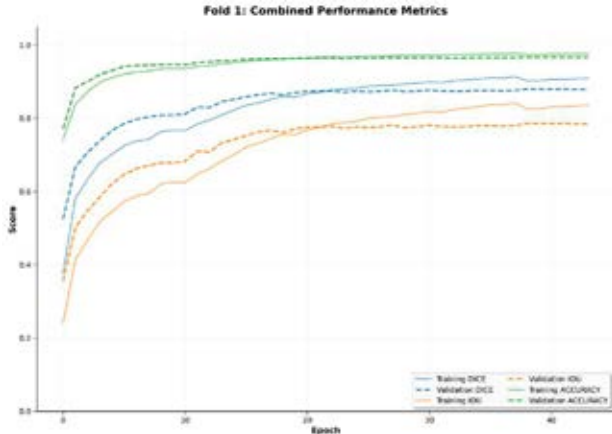


Fig. 2. Fold 1 the combined learning curves (Dice, IoU, Accuracy). Solid lines: training; dashed: validation.

connected component. The model with highest validation Dice is saved.

Fig. 3 illustrates the impact of our morphological post-processing. The bottom row particularly demonstrates how this approach handles fragmented predictions: the raw output shows two disconnected regions, but the morphological operations eliminate the spurious smaller fragment while preserving the main nodule region. However, as illustrated in Fig. 4, morphology can also slightly shrink to thin or highly under-segmented regions, especially in very small nodules, and may occasionally eliminate true detections if the nodule is fragmented into multiple tiny islands.

B. Ablation study

To systematically evaluate the contribution of each proposed component, we conducted a comprehensive ablation study by progressively adding modules to a baseline UNet architecture. Table II shows the combination of DPAM and adaptive threshold search achieves the best performance with a Dice coefficient of 0.8875, representing a 2.5% relative improvement over the baseline. The negative impact of morphological post-processing underscores the importance of careful component selection, as additional processing steps do not always guarantee performance improvements in medical imaging applications. DPAM provides the largest single contribution highlighting the importance of attention mechanisms for medical image segmentation tasks involving complex anatomical structures.

C. Comparative Experiment

To validate the effectiveness of our proposed architecture we conducted a comparative analysis against several thyroid segmentation methods. Our model achieved superior

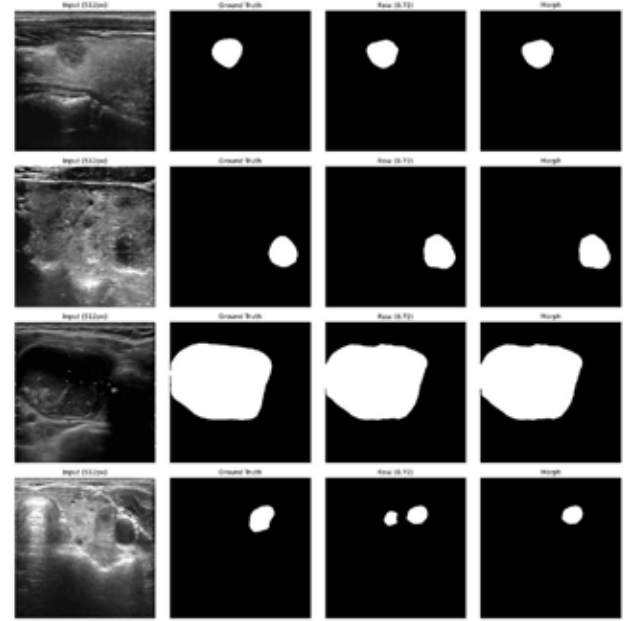


Fig. 3. Validation examples from Fold 1 comparing raw predictions with morphological post-processing. Columns show (left to right): input ultrasound image, ground truth mask, raw model output, and result after morphological post-processing.

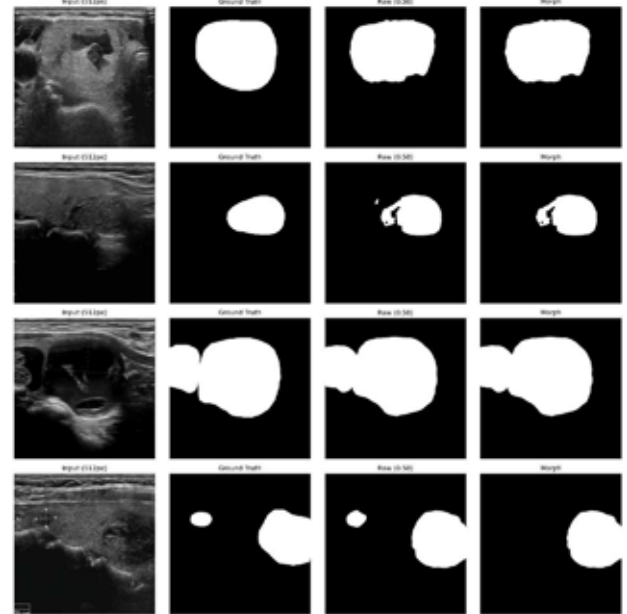


Fig. 4. Validation examples from Fold 2 comparing raw predictions with morphological post-processing. Columns show (left to right): input ultrasound image, ground truth mask, raw model output, and result after morphological post-processing

performance, demonstrating the highest Dice (0.8875), IoU (0.7969), Recall (0.8827), and Precision (0.8955) compared to all other tested models. Notably, the improvement over standard approaches like UNet and more recent enhancements such as BPAT-UNet, DPAM-UNet++ and Improved V-Net highlights the robustness and effectiveness of integrating

Table 2. Ablation study results

	Dice	IoU	Acc	Recall	Precision
Baseline	0.8658	0.7691	0.9620	0.8700	0.8750
DPAM	0.8853	0.7931	0.9688	0.8775	0.8956
DPAM + Threshold	0.8875	0.7969	0.9688	0.8827	0.8955
DPAM + Threshold + Morph	0.8836	0.7914	0.9681	0.8770	0.8966

Table 3. Performance comparison of different models

	DICE	IoU	Recall	Precision
Proposed Model	0.8875	0.7969	0.8827	0.8955
BPAT-UNet [9]	0.8423	0.7187	0.8557	0.8294
DPAM-UNet++ [10]	0.8310	0.7451	0.8702	0.8443
Improved V-Net [12]	0.8338	0.7550	-	-
UNet [4]	0.7829	0.6395		
SGUNet [2]	0.7323	0.6298	0.7120	0.8522

EfficientNet-B5, DPAM, and composite loss functions specifically tailored for thyroid ultrasound image segmentation.

IV. CONCLUSION

In this study, we introduced an EfficientNet-B5-based UNet++ segmentation architecture augmented with Dual-Path Attention Modules (DPAM), optimized for automated thyroid nodule segmentation in ultrasound images. The proposed model demonstrates strong performance, effectively addressing challenges such as low contrast, speckle noise, and blurred boundaries through dual-path attention mechanisms and composite loss optimization. Experimental results validate significant improvements over existing segmentation methods, indicating the robustness and clinical applicability of our approach.

However, our study is limited by the size and diversity of the dataset, suggesting potential improvements through training on larger and multi-center data. Future research directions include exploring advanced domain adaptation techniques and integrating explainability modules, further bridging the gap between automated segmentation and clinical practice.

REFERENCES

- [1] A. Prochazka and J. Zeman, "Thyroid Nodule Segmentation in Ultrasound Images Using U-Net With ResNet Encoder: Achieving State-of-the-Art Performance on All Public Datasets," *AIMS Med. Sci.*, vol. 12, no. 2, pp. 124–144, Apr. 2025.
- [2] H. Pan, Q. Zhou, and L. J. Latecki, "SGUNet: Semantic Guided UNet for Thyroid Nodule Segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, Venice, Italy, Apr. 2020, pp. 72–75.
- [3] J. Sun et al., "TNSNet: Thyroid Nodule Segmentation in Ultrasound Imaging Using Soft Shape Supervision," *Comput. Methods Programs Biomed.*, vol. 215, Art. no. 106600, Aug. 2022.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.
- [5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "U-Net++: Redesigning Skip Connections to Exploit Multi-Scale Features in Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
- [6] H. Gong et al., "Thyroid Region Prior-Guided Attention for Ultrasound Segmentation of Thyroid Nodules," *Comput. Biol. Med.*, vol. 155, Art. no. 106389, Mar. 2023.
- [7] A. Ozcan, Ö. Tosun, E. Donmez, and M. Sanwal, "Enhanced-TransUNet for Ultrasound Segmentation of Thyroid Nodules," *Biomed. Signal Process. Control*, vol. 95, Art. no. 106472, 2024.
- [8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, May 2019, pp. 6105–6114.
- [9] H. Bi et al., "BPAT-UNet: Boundary Preserving Assembled Transformer UNet for Ultrasound Thyroid Nodule Segmentation," *Comput. Methods Programs Biomed.*, vol. 238, Art. no. 107614, Jan. 2023.
- [10] P. Dong et al., "An Ultrasound Image Segmentation Method for Thyroid Nodules Based on Dual-Path Attention Mechanism-Enhanced U-Net++," *BMC Med. Imaging*, vol. 24, Art. no. 341, 2024.
- [11] H. Kervadec et al., "Boundary Loss for Highly Unbalanced Segmentation," in *Proc. Med. Imaging with Deep Learn. (MIDL)*, Innsbruck, Austria, July 2019, pp. 285–296.
- [12] B. Yetginler and İ. Atacak, "An Improved V-Net Model for Thyroid Nodule Segmentation," *Appl. Sci.*, vol. 15, no. 7, Art. no. 3873, 2025.
- [13] N. Abraham and N. M. Khan, "A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imaging (ISBI)*, Venice, Italy, Apr. 2019, pp. 672–676.
- [14] H. Gong et al., "TN3K: Thyroid Nodule 3493 A Public Dataset for Thyroid Nodule Segmentation," *Comput. Methods Programs Biomed.*, vol. 215, Art. no. 106613, 2022.
- [15] Y. Cao et al., "GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Seoul, Korea, Oct. 2019, pp. 1971–1980.
- [16] Q. Wang et al., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, June 2020, pp. 11534–11542.
- [17] S. Woo et al., "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sept. 2018, pp. 3–19.