GAT_ASP-UNet: Unified Deep Learning Approaches — Ensemble-Based U-Net for Medical Image Segmentation

RAHUL DRABIT CHOWDHURY*, Bangladesh University of Engineering and Technology, Bangladesh
MASAB HASNAIN†, Bangladesh University of Engineering and Technology, Bangladesh
MAREFUL ISLAM‡, Bangladesh University of Engineering and Technology, Bangladesh

Medical image segmentation demands precise boundary delineation even under dense-distribution and fuzzy-edge conditions. We present **GAT_ASP-UNet**, a novel encoder–decoder architecture that couples a Fixed-Volume Compressor (FSCM) with a Graph Attention (GAT) Bridge to introduce structured relational reasoning into skip connections at constant complexity $O(1024)$, independent of input resolution. Evaluated on four public benchmarks (CVC-ClinicDB, Kvasir-SEG, ISIC2018, and Breast Ultrasound-B), the model achieves a Dice score of 90.88% on dermoscopic data and demonstrates that GAT-enhanced skip connections improve boundary recall over standard baselines.

Additional Key Words and Phrases: medical image segmentation, U-Net, graph attention network, fixed-volume compression, ADFM, polyp detection

1 Introduction

Medical image segmentation is a critical prerequisite for objective clinical diagnosis, providing the spatial foundations for condition evaluation and surgical planning. Despite the dominance of U-shaped convolutional neural networks (CNNs), significant challenges persist in the precise delineation of "fuzzy edges" and the separation of objects within "dense distributions" [1, 2]. These difficulties arise from the inherent local inductive bias of standard convolutions, which often fail to capture long-range contextual dependencies, and the prohibitive computational costs of Vision Transformers (ViTs) when processing high-resolution medical volumes.

The motivation for this research is to enhance the "reusability of features" and "contextual reasoning" within a resource-constrained framework. Existing architectures such as MFA U-Net attempt multi-stage extraction, while DDS-UNet focuses on mitigating semantic gaps, yet neither fully addresses the need for structured relational reasoning across skip connections.

This paper introduces **GAT_ASP-UNet**, which leverages a novel integration of the Fixed Volume Compression module (FSCM) and a Graph Attention (GAT) Bridge. By constraining intermediate features to a fixed 32×32 volume, the model ensures that the complexity of relational reasoning is a constant factor $O(1024)$, independent of input resolution.

2 Related Works

The development of medical imaging architectures has focused on refining the encoder–decoder paradigm to bridge the gap between low-level spatial detail and high-level semantics.

*Group 6 | Student ID: 0424057003 | Mobile: 01858017317
†Group 6 | Student ID: 0424052099 | Mobile: 01534306001
‡Group 6 | Student ID: 0424056005 | Mobile: 01746390162

Authors' Contact Information: Rahul Drabit Chowdhury, Bangladesh University of Engineering and Technology, Department of Computer Science and Engineering, Bangladesh, rahuldrabit@gmail.com; Masab Hasnain, Bangladesh University of Engineering and Technology, Department of Computer Science and Engineering, Bangladesh, masabhasnain1@gmail.com; Mareful Islam, Bangladesh University of Engineering and Technology, Department of Computer Science and Engineering, Bangladesh, 0424056005@grad.cse.buet.ac.bd.

2.1 Standard U-Net

The baseline U-Net architecture utilises a symmetric contracting path to capture context and a symmetric expansive path to enable precise localisation via skip connections [1].

2.2 DDS-UNet

The Deep Dynamic Self-Adjusting U-Net (DDS-UNet) introduces the Lightweight Deformable Residual (LDR) module and a Semantic Mitigation Module (SMM). The LDR enhances the receptive field for contour extraction, while the SMM reduces the "semantic gap" between codec layers—discrepancies that inherently increase with network depth [2].

2.3 MFA U-Net

The Multi-Stage Feature Analysis Network (MFA U-Net) emphasises "multi-channel dimensional feature extraction" by reusing input images across parallel convolutional branches. Its primary extraction block processes inputs through channels of 8, 16, 32, and 64, fused into a 120-channel feature map before reduction back to 64 channels [3].

2.4 DoubleU-NetPlus

DoubleU-NetPlus extends the dual U-Net paradigm with a context-guided attention mechanism and multi-scale residual feature fusion. By coupling two U-Net branches with cross-attention, it captures both local boundary detail and global semantic context for robust semantic segmentation of medical images [4].

2.5 UCTransNet

UCTransNet replaces the standard skip connections in U-Net with a Channel-wise Cross-Fusion Transformer (CCT) module. This channel-wise perspective enables selective multi-scale feature fusion between encoder and decoder stages, substantially reducing the semantic gap inherent in deep skip connections [5].

3 Proposed Model

The GAT_ASP-UNet is engineered to provide multi-scale feature enrichment and relational context through a dual-path encoder and a triple-concatenation decoder. The overall architecture is illustrated in Figure 1 (see Appendix).

3.1 Advanced Feature Module (ADFM)

The ADFM is the primary engine for multi-scale context capture (Figure 2a). It employs parallel branches with 1×1, 3×3, 5×5, and dilated kernels alongside an ASPP block. Each branch incorporates a "downsampled self-attention" mechanism that operates on a reduced grid to manage computational overhead. Concatenated branch outputs are passed through a global multi-head attention (MHA) layer before final 1×1 fusion, with a residual connection preserving input information.

3.2 Fixed Volume Compressor (FSCM)

To ensure computational stability, the FSCM forces features into a compact 32×32 spatial volume using 1×1 channel reduction followed by adaptive average pooling. This bounds the complexity of subsequent attention and GAT layers to a constant factor $O(1024)$, making the most intensive parts of the network resolution-independent. The FSCM pipeline is shown in Figure 3 (see Appendix).

3.3 Graph Attention (GAT) Bridge

The GAT Bridge converts spatial features into a grid graph to perform relational reasoning (Figure 2b). Features are reduced to a target channel dimension C_r and pooled to a $P \times P$ grid. A 4-neighbourhood grid graph is constructed and a PyG `GATConv` layer is applied over the nodes. Following graph processing, features are upsampled and restored via a 1×1 convolution to match the skip connection dimensions.

3.4 GhostRFBCoordBottleneck

This lightweight bottleneck (Figure 4) utilises *GhostModules* for cheap feature expansion, a *LiteRFB* module (multi-branch dilated depthwise convolutions) to expand the receptive field, and *CoordinateAttention* to refine features through spatial-aware channel gating (Figure 4, Appendix).

3.5 Architectural Rationale

The dual-path encoder at each stage uses:

- **Path A:** encoder output → FSCM → ADFM → channel adapt → upsample to fuse.
- **Path B:** encoder output → MaxPool/downsample.

Both paths are concatenated and fused via a 1×1 convolution. In the decoder, a "triple-concat" strategy merges:

(1) direct encoder features,
(2) GAT bridge outputs, and
(3) previous decoder features re-processed via FSCM–ADFM.

4 Experiments and Results

4.1 Experimental Setup

- **Dataset split:** 80% training, 20% validation.
- **Training:** 100 epochs, learning rate $\eta = 10^{-4}$, Adam optimiser.
- **Loss Functions:** Dice + BCE (primary); Focal Tversky + IoU (secondary).
- **Augmentation:** 9 techniques including elastic transform, colour jitter, and random occlusion.
- **Metrics:** IoU, Dice, Precision, Recall, Accuracy.

4.2 Datasets

Table 1. Datasets used for evaluation.

Dataset	Modality / Target	Images
CVC-ClinicDB	Endoscopic / Polyp detection	612
ISIC2018	Dermoscopic / Skin lesions	2,596
Kvasir-SEG	Endoscopic / Polyp images	1,000
Breast Ultrasound B	Ultrasound / Breast lesions	163

4.3 Performance on External Datasets

The model generalises exceptionally well to dermoscopic images (ISIC2018 Dice: 90.88%). The lower IoU/Dice on Breast Ultrasound B suggests a need for domain-specific augmentation to handle ultrasound noise profiles. Figure 5 shows qualitative predictions across datasets.

Table 2. Results on external datasets (all values in %).

Dataset	Split	IoU	Dice	Prec	Rec	Acc
Kvasir-SEG	Val	76.17	86.14	88.99	84.27	95.86
	Train	74.48	85.03	88.07	83.13	95.60
ISIC2018	Val	83.61	90.88	93.64	88.81	96.24
	Train	82.62	90.28	92.37	88.88	96.04
Breast US B	Val	66.64	79.40	82.75	76.88	98.17
	Train	63.54	77.37	73.91	83.34	97.86

4.4 Comparative Analysis on CVC-ClinicDB

Table 3. Comparative analysis on CVC-ClinicDB (metrics in %).

Model / Variant	Val Loss	Val IoU	Val Dice	Val Acc	Val Prec	Val Rec
DDSUNet – Dice loss	0.11762	86.74	92.63	98.61	94.69	91.33
DDSUNet – Dice + BCE	0.04925	85.28	91.77	98.51	93.77	90.53
DDSUNet + GATConv	0.12696	86.39	92.47	98.60	93.77	91.91
Proposed – Focal Tversky	0.21342	72.58	83.22	96.76	83.59	84.99
Proposed – Dice + BCE	0.23182	76.30	86.14	97.54	88.98	84.58

While the proposed model demonstrates strong training metrics (Train Dice ≈90.9% on CVC), its validation performance lags behind the DDSUNet baselines, indicating overfitting or domain mismatch. Notably, the GATConv baseline variant achieved the highest validation recall (91.91%), empirically supporting the thesis that graph-based relational reasoning captures complex non-local boundaries.

5 Future Works and Conclusion

To mitigate the observed overfitting and improve generalisation, future work should prioritise:

- **Enhanced Regularisation:** Dropout and weight decay within ADFM and GAT Bridge for robustness to domain shifts.
- **Structural Ablations:** Investigation of GAT layer depth and attention head counts to optimise the balance between relational reasoning and parameter count.
- **Data Augmentation:** Stronger elastic transforms and random occlusion to better simulate variability in clinical endoscopic imaging.

In conclusion, GAT_ASP-UNet provides a sophisticated framework for integrating relational reasoning into medical segmentation. While it shows particular strength in dermoscopic analysis, further tuning is required to match state-of-the-art performance on endoscopic datasets.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015, pp. 234–241.

[2] Y. Ou, Y. Chen, Y. Zhang, S. Yang, X. Zhang, Y. Zhou, S. Chen, and L. Peng, "Enhanced medical image segmentation via deep dynamic self-adjusting U-Net with multi-scale attention and semantic mitigation," *The Visual Computer*, vol. 41, pp. 8385–8401, 2025.

[3] Y. Wang, S. Wang, and J. He, "MFA U-Net: a U-Net like multi-stage feature analysis network for medical image segmentation," *Pattern Analysis and Applications*, vol. 27, p. 110, 2024.

[4] M. R. Ahmed, A. F. Ashrafi, R. U. Ahmed, S. Shatabda, A. K. M. M. Islam, and S. Islam, "DoubleU-NetPlus: a novel attention and context-guided dual U-Net with multi-scale residual feature fusion network for semantic segmentation of medical images," *Neural Computing and Applications*, vol. 35, pp. 14379–14401, 2023.

[5] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI*, vol. 36, no. 3, 2022, pp. 2441–2449.

A Architecture Diagrams and Results

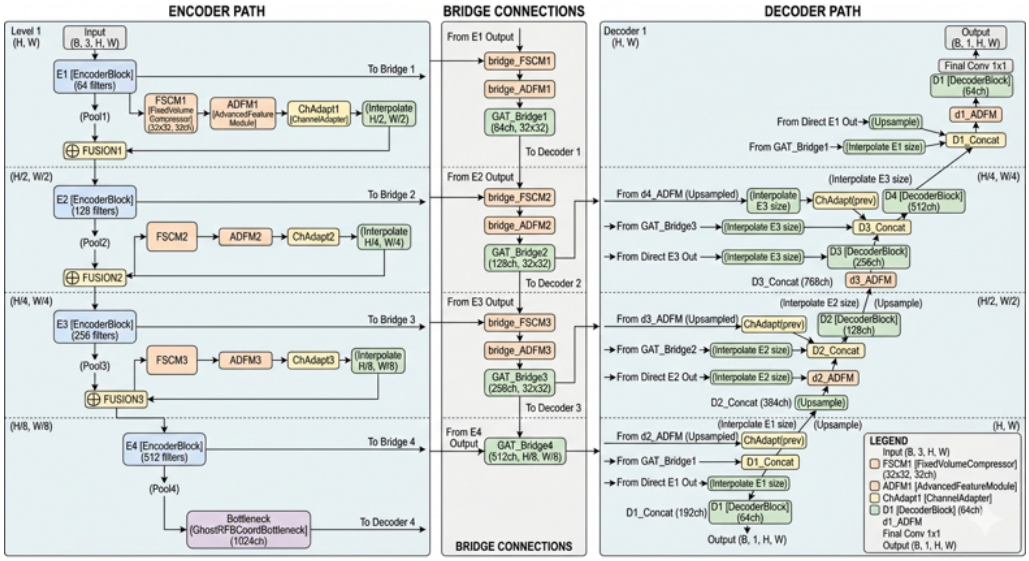
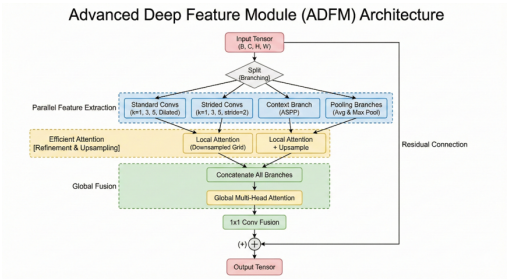
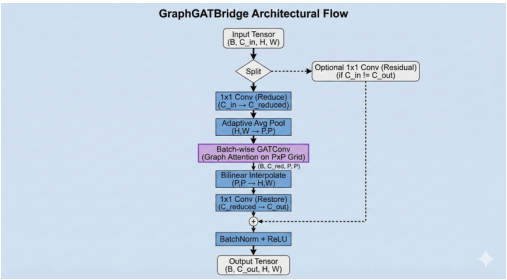


Fig. 1. Overall architecture of GAT_ASP-UNet. The dual-path encoder (Path A: FSCM→ADFM; Path B: MaxPool) feeds enriched features to the GAT Bridge skip connections. The triple-concat decoder (direct encoder | GAT skip | re-processed prev. decoder) reconstructs the segmentation mask.



(a) ADFM module.



(b) GAT Bridge module.

Fig. 2. Key sub-modules. (a) ADFM: 10 parallel branches with per-branch downsampled self-attention followed by global MHA. (b) GAT Bridge: pools to a $P \times P$ grid graph and applies GATConv before restoring spatial resolution.

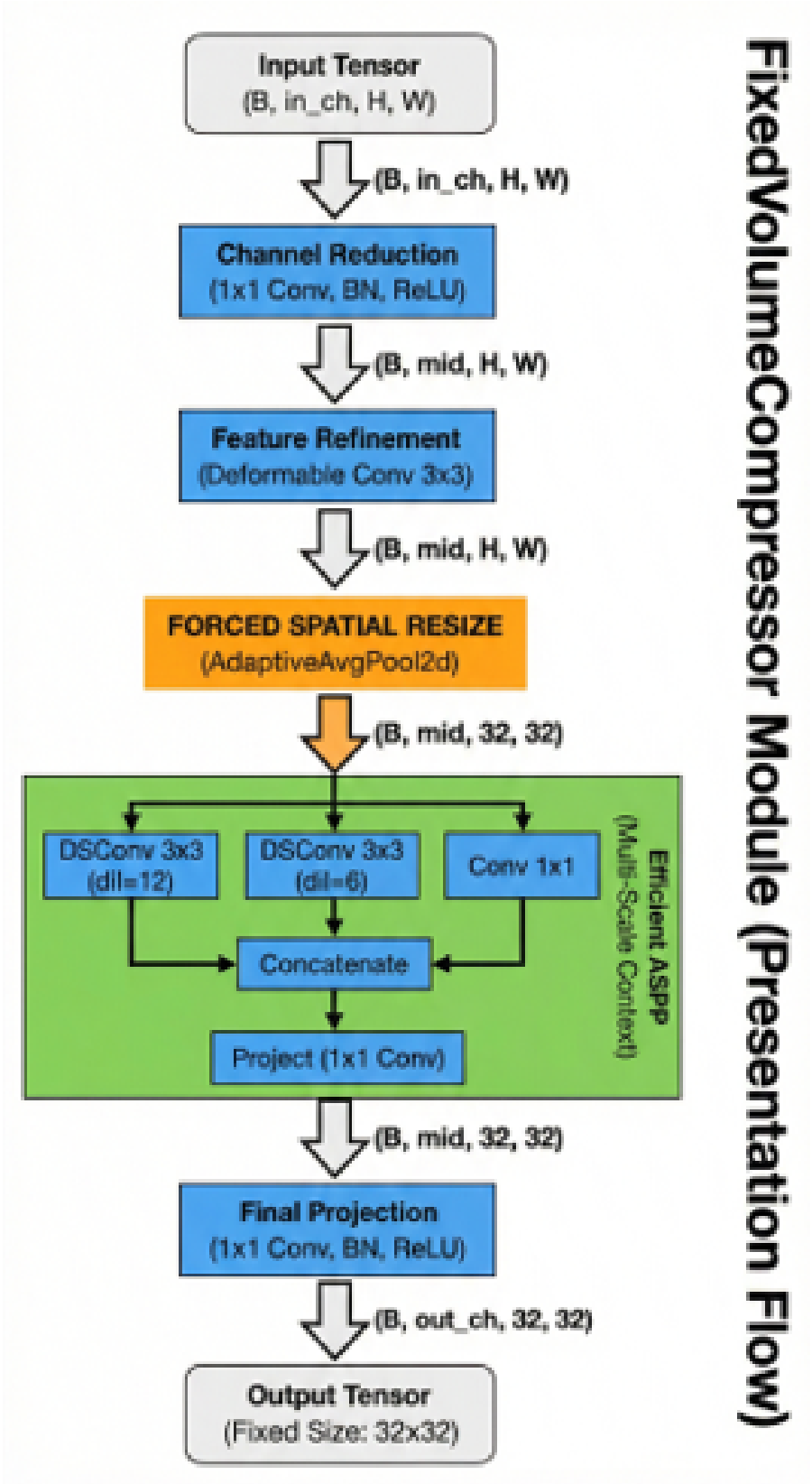


Fig. 3. Fixed Volume Compressor (FSCM): 1×1 channel reduction → deformable conv → AdaptiveAvgPool to

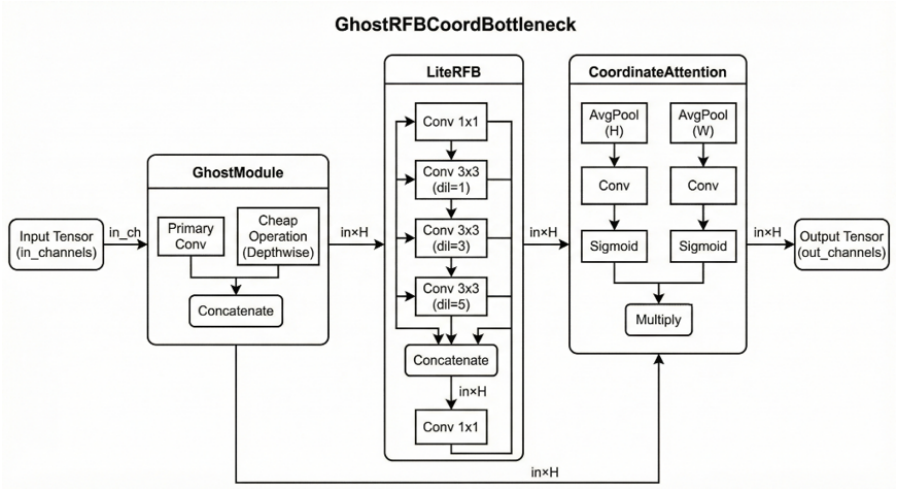


Fig. 4. GhostRFBCoordBottleneck: GhostModule expansion → LiteRFB multi-dilated depthwise branches → CoordinateAttention channel-spatial gating.

Fig. 5. Qualitative segmentation results on CVC-ClinicDB from the final GAT_ASP-UNet model. Each row shows the original endoscopic image, the ground-truth mask, and the predicted mask.

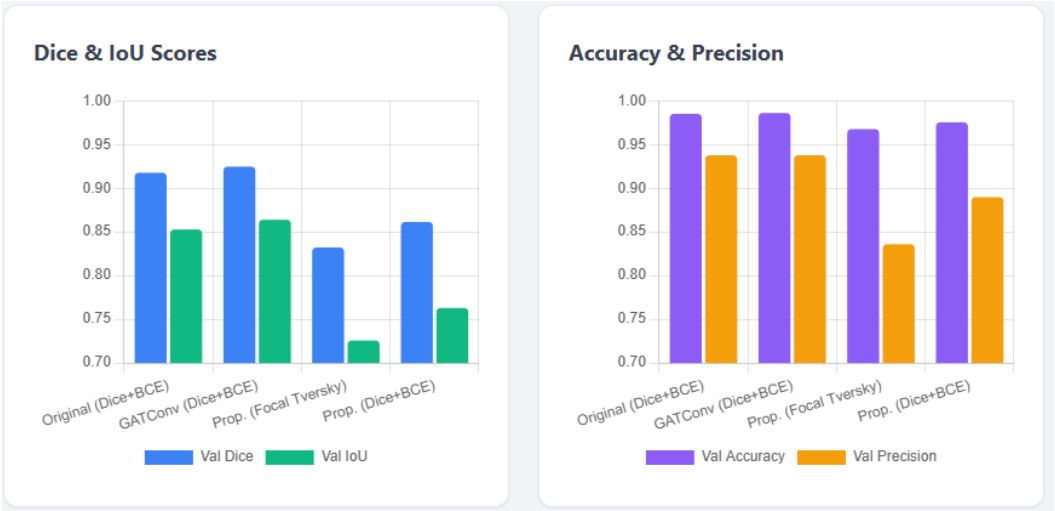


Fig. 6. Comparative results across baseline models and the proposed GAT_ASP-UNet. Columns represent different architectures; rows correspond to evaluation metrics on CVC-ClinicDB.