2025 28th International Conference on Computer and Information Technology (ICCIT)
19-21 December 2025, Cox's Bazar, Bangladesh

SwinTextUNet: Integrating CLIP-Based Text Guidance into Swin Transformer U-Nets for Medical Image Segmentation

Ashfak Yeafi
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
Email: yeafiashfak@gmail.com

Parthaw Goswami
Department of Electronics and Communication Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
Email: parthawgoswami555@gmail.com

Md Khairul Islam
Department of Mathematics and Computer Science
Hobart and William Smith Colleges
Geneva, NY, USA
Email: khairul.robotics@gmail.com

Ashifa Islam Shamme
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
Email: ashifa54islam@gmail.com

2025 28th International Conference on Computer and Information Technology (ICCIT) | 979-8-3315-7867-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCIT68739.2025.11491606

***Abstract*—Precise medical image segmentation is fundamental for enabling computer-aided diagnosis and effective treatment planning. Traditional models that rely solely on visual features often struggle when confronted with ambiguous or low-contrast patterns. To overcome these limitations, we introduce SwinTextUNet, a multimodal segmentation framework that incorporates Contrastive Language-Image Pre-training (CLIP), derived textual embeddings into a Swin Transformer U-Net backbone. By integrating cross-attention and convolutional fusion, the model effectively aligns semantic text guidance with hierarchical visual representations, enhancing robustness and accuracy. We evaluate our approach on the QaTa-COV19 dataset, where the proposed four-stage variant achieves an optimal balance between performance and complexity, yielding Dice and IoU scores of 86.47% and 78.2%, respectively. Ablation studies further validate the importance of text guidance and multimodal fusion. These findings underscore the promise of vision–language integration in advancing medical image segmentation and supporting clinically meaningful diagnostic tools.**

***Index Terms*—Vision–Language Models, CLIP, Swin Transformer, Cross-Attention, Medical Image Segmentation**

I. Introduction

Precise segmentation of medical images is crucial for computer-assisted diagnosis (CAD), disease measurement, and treatment planning. Classical CNN-based models, such as U-Net and its extensions, have achieved strong results attributed to their encoder–decoder architecture incorporating skip connections. However, these frameworks rely solely on visual features, ignoring textual information (e.g., radiology reports or diagnostic notes) that clinicians routinely use to interpret images. This limitation reduces robustness when image features are ambiguous or noisy. Recent advances in transformer architectures, particularly the Swin Transformer [1], have enabled long-range dependency modeling with hierarchical efficiency, outperforming CNNs in both natural and medical image segmentation. Parallelly, vision–language models (VLMs) such as CLIP [2] have demonstrated remarkable cross-modal alignment by jointly training image and text encoders. Despite their success in natural image domains, CLIP's integration into medical segmentation remains underexplored. In order to fill this void, we suggest SwinTextUNet, a multimodal segmentation framework that embeds CLIP-derived textual features into a Swin Transformer U-Net backbone. By fusing semantic text cues with visual features, our model improves segmentation robustness and captures clinically meaningful context. The key contributions of this work can be summarized as follows:

1) A Swin Transformer U-Net variant that integrates CLIP-based textual embeddings for multimodal medical segmentation.
2) A text-conditioning mechanism that fuses semantic embeddings with image tokens across encoder and decoder stages for enhanced cross-modal alignment.

This paper's is structured as follows: The work is reviewed in Section II, the methodology is described in Section III, the experimental results are reported in Section IV, and the work is concluded in Section V.

II. Related work

Medical image segmentation has evolved through multiple architectural paradigms, ranging from convolutional encoder–decoders to transformer-based models and, more recently, multimodal vision–language approaches. In this section, we review these developments, with a particular focus on biomedical segmentation networks that are closely related to our proposed framework.

A. CNN-Based Architectures

The primary model for medical segmentation is still the U-Net [3], which uses an encoder–decoder with skip links to maintain spatial detail. Numerous variants have extended

this baseline: UNet++ [4] redesigned skip pathways for improved semantic fusion. Specialized CNN extensions have been proposed for specific clinical domains. For example, ADTNet [5] introduced an Attention-Guided U-Net with Dynamic Convolution and Transformers for skin cancer segmentation. Similarly, GSNet [6] and DSNet [7] leveraged 3D convolutions and attention-based skip connections for glioma segmentation in the BraTS benchmarks. These works illustrate the potential of augmenting U-Net structures with domain-specific modules for improved performance in challenging segmentation tasks [8]–[10].

B. Transformer-Based Architectures

Transformers [11] have provided a powerful alternative to CNNs by capturing global context through self-attention. Their adaptation to medical segmentation has been highly impactful. TransUNet [12] combined CNN backbones with ViT encoders to balance local and global feature extraction. Swin-UNet [13], based on hierarchical Swin Transformers [1], introduced shifted window attention for scalable context modeling across resolutions. Further variants such as MISSFormer [14], MedT [15] refined hierarchical attention, token aggregation, and gated mechanisms to enhance boundary delineation. Model like EF-SwinNet [16] uses a hybrid architecture combining CNN and transformer for medical image analysis. While these architectures significantly outperform only CNN-based models, they are still limited to unimodal (image-only) input.

C. Vision–Language Segmentation

The emergence of large-scale contrastive pretraining has sparked interest in multimodal segmentation [17]. CLIP [2] demonstrated powerful alignment between vision and text embeddings. Medical adaptations such as BioViL [18], PubMedCLIP [19] extended these models to radiology data, enabling classification, retrieval, and report understanding. For segmentation, recent works integrated CLIP with pixel-level architectures, MedCLIP-SAM [20] fused CLIP features with the Segment Anything Model for general-purpose medical segmentation. Although promising, these models often rely on external foundation modules or operate in weakly supervised settings [21], limiting their applicability in specialized domains. In contrast, our proposed SwinTextUNet introduces a fully supervised, end-to-end multimodal segmentation framework. By fusing CLIP-based text embeddings with hierarchical Swin Transformer U-Net features through cross-attention and ConvFuse blocks, our approach explicitly aligns textual priors with multiscale visual features. Unlike prior CNN-based or transformer-based methods, SwinTextUNet leverages both modalities to enhance segmentation quality, representing a novel contribution in medical vision–language modeling.

III. Methodology

We propose SwinTextUNet, a multimodal segmentation framework that integrates CLIP-based textual guidance into a hierarchical Swin Transformer U-Net. The model is designed to capture multi-scale visual features while incorporating semantic priors from domain-specific medical text. Four primary components make up the architecture, as seen in Figure 4: (i) CLIP-based text encoding, (ii) a hierarchical Swin Transformer encoder, (iii) text-guided cross-attention and convolutional fusion modules, and (iv) a decoder that reconstructs full-resolution segmentation masks. Key building blocks are further highlighted in Figures 1 (Swin Transformer Block), 2 (Cross-Attention Block), and 3 (ConvFuse Block).

A. Text Encoding via CLIP

To incorporate semantic guidance, we employ a pretrained CLIP text encoder [2]. Given a batch of B text prompts

$$\mathscr{T} = \{t^{(b)}\}_{b=1}^{B},$$

the encoder outputs pooled embeddings $\bar{Z}_t \in \mathbb{R}^{B\times D_t}$. Since OpenAI CLIP returns a single pooled vector, we form a token sequence by unsqueezing it, i.e., $Z_t \in \mathbb{R}^{B\times 1\times D_t}$. A linear projection aligns dimensions:

$$\tilde{Z}_t = Z_t W_t, \quad \tilde{\bar{Z}}_t = \bar{Z}_t W_t, \quad W_t \in \mathbb{R}^{D_t\times D_v}, \tag{1}$$

where D_v is the visual token dimension. The text tower remains frozen to preserve semantic priors, while W_t is optimized during training.

B. Visual Encoding with Swin Transformer

The visual encoder is built upon the hierarchical Swin Transformer [1]. The input image $X \in \mathbb{R}^{B\times 3\times H\times W}$ is first divided into non-overlapping grous of shape $P\times P$ and projected into tokens:

$$Z_1 \in \mathbb{R}^{B\times N_1\times C_1}, \quad N_1 = \tfrac{H}{P}\cdot\tfrac{W}{P}, C_1 = 96, \quad (P=4). \tag{2}$$

At each subsequent stage, the spatial resolution is halved while the channel dimension is doubled:

$$\begin{aligned} Z_1 &\in \mathbb{R}^{B\times(56^2)\times 96}, \\ Z_2 &\in \mathbb{R}^{B\times(28^2)\times 192}, \\ Z_3 &\in \mathbb{R}^{B\times(14^2)\times 384}, \\ Z_4 &\in \mathbb{R}^{B\times(7^2)\times 768}, \end{aligned} \tag{3}$$

for input resolution 224×224. The model can capture both global semantic context and fine-grained information attributable to its hierarchical nature. For decoding, stage outputs are reshaped into feature maps $S_s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$, which are later used as skip connections.

C. Shifted-Window Self-Attention

Each Swin block applies self-attention within local $M\times M$ windows:

$$\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_{\mathrm{rel}}\right)V, \tag{4}$$

where B_{rel} is the relative position bias. A cyclic shift of $(M/2, M/2)$ between consecutive blocks enables inter-window connections, while a precomputed attention mask prevents invalid interactions. This reduces computational complexity from global $\mathscr{O}((HW)^2)$ to local $\mathscr{O}(HW\cdot M^2)$, enabling efficient high-resolution processing.

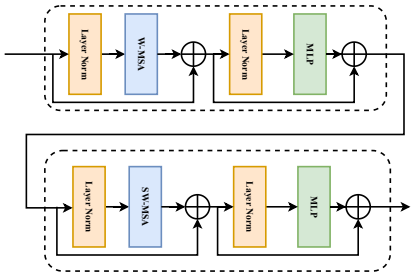

Fig. 1. Illustration of the Swin Transformer block with windowed and shifted window attention.

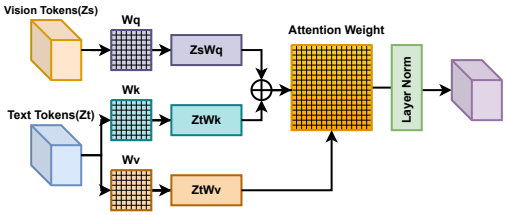

Fig. 2. Cross-Attention block. Vision tokens attend to text tokens, refining representations with semantic guidance.

D. Text-Guided Cross-Attention

To integrate semantics, we refine vision tokens with text tokens at each stage s. Let $Z_s \in \mathbb{R}^{B\times N_s\times C_s}$ and $\tilde{Z}_t \in \mathbb{R}^{B\times T\times C_s}$ with $T=1$. Multi-head cross-attention computes:

$$Q = Z_s W_Q, \quad K = \tilde{Z}_t W_K, \quad V = \tilde{Z}_t W_V, \tag{5}$$

$$\tilde{Z}_s = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{6}$$

$$Z'_s = \mathrm{LN}\big(Z_s + \tilde{Z}_s\big), \quad Z''_s = Z'_s + \mathrm{MLP}(\mathrm{LN}(Z'_s)). \tag{7}$$

This asymmetric design ensures that textual context steers visual features without overwhelming local structures. Complexity is $\mathcal{O}(BN_sTC_s)$, negligible since $T \ll N_s$.

E. Convolutional Fusion (ConvFuse)

The ConvFuse block merges upsampled decoder features U_s with skip maps S_s via convolutional refinement:

$$F_s = \phi\big([S_s, U_s]\big), \tag{8}$$

where $[\cdot]$ denotes channel concatenation and ϕ consists of two 3×3 convolutions. Unlike pure token fusion, ConvFuse exploits local spatial continuity and improves boundary reconstruction. At the bottleneck, we additionally fuse S_4 into Z_4 to enrich deep features with fine-grained context.

F. Decoder and Reconstruction

Decoding proceeds via progressive PatchExpand followed by ConvFuse:

$$\hat{Z}_4 \to Y_3 \to Y_2 \to Y_1 \to Y_0,$$

with resolutions $\{7^2, 14^2, 28^2, 56^2, 112^2\}$. Tokens are converted to maps at each step, concatenated with skips, fused, and re-tokenized. The final map $Y_0 \in \mathbb{R}^{B\times C_0/2\times 112\times 112}$ is upsampled to 224×224 and processed through a 1×1 convolution to produce output.

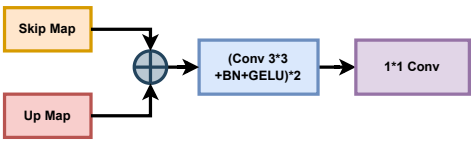

Fig. 3. ConvFuse block. Skip map and upsampled tokens are concatenated, refined by convolutions, and re-tokenized.

G. Output and Loss Function

For binary segmentation, the output applies a sigmoid:

$$\hat{Y} = \sigma(\hat{Y}_{\text{logits}}), \quad \sigma(x) = \frac{1}{1+e^{-x}}. \tag{9}$$

We optimize a hybrid loss that combines cross-entropy and dice:

$$\mathscr{L} = \lambda_{\text{dice}}\mathscr{L}_{\text{Dice}} + \lambda_{\text{ce}}\mathscr{L}_{\text{CE}}, \tag{10}$$

with

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\big], \tag{11}$$

$$\mathscr{L}_{\text{Dice}} = 1 - \frac{2\sum_i y_i\hat{y}_i + \varepsilon}{\sum_i y_i + \sum_i \hat{y}_i + \varepsilon}, \ \varepsilon = 10^{-6}. \tag{12}$$

This balances overlap with per-pixel classification, alleviating class imbalance in medical segmentation.

H. Training Pipeline

Algorithm 1 summarizes the complete training pipeline of SwinTextUNet.

IV. EXPERIMENT AND RESULT

A. Dataset Description

We evaluate the proposed SwinTextUNet on the QaTa-COV19 dataset [22], a large-scale benchmark of 9,258 chest X-ray (CXR) radiographs with manually annotated COVID-19 lesions. Each image is paired with a binary lesion mask delineating infected lung regions, enabling robust segmentation evaluation. In addition, the dataset includes textual annotations curated by medical experts [23], describing infection patterns such as lesion count, anatomical location, and laterality (unilateral vs. bilateral). For example, *"Bilateral pulmonary infection, two infected areas, upper left lung and upper right lung"*. These annotations are used as input to the CLIP text encoder, providing semantic priors to guide segmentation. Figure 5 illustrates sample triplets of CXR, mask, and annotation across subsets.

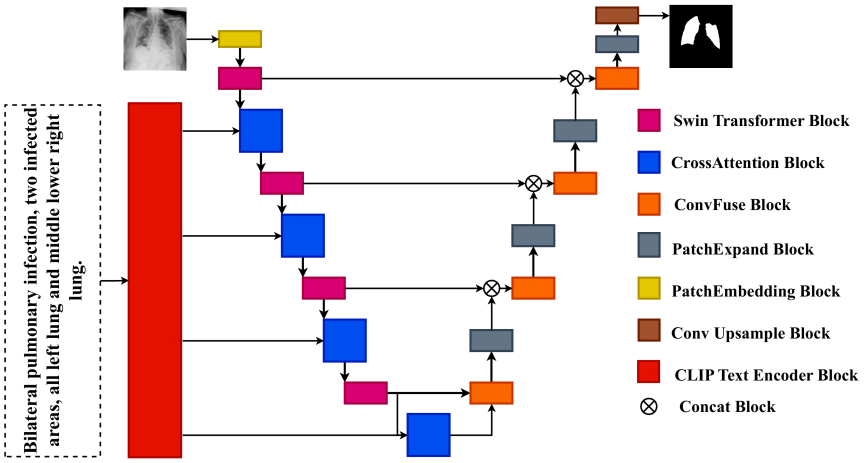

Fig. 4. Overall architecture of SwinTextUNet. The model integrates CLIP-based text embeddings with a hierarchical Swin U-Net through cross-attention and ConvFuse blocks.

Algorithm 1: Training pipeline of SwinTextUNet

Input: Image $X \in \mathbb{R}^{B\times 3\times H\times W}$, text prompts $\mathscr{T}$, ground-truth mask Y.
Output: Predicted segmentation $\hat{Y}$.

Step 1: Text Encoding
Encode $\mathscr{T}$ with CLIP: $(Z_t, \bar{Z}_t) = f_{\text{text}}(\mathscr{T})$.
Project to visual dimension: $\tilde{Z}_t = W_t Z_t$.
Step 2: Image Encoding
Patch embed image X: $Z_0 = f_{\text{patch}}(X)$.
Run Swin Transformer blocks at each stage $s = 1..4$ to obtain Z_s.
Extract skip maps S_s for decoder.
Step 3: Cross-Attention Fusion
For each stage s, update tokens via $\hat{Z}_s = \text{CrossAttn}(Z_s, \tilde{Z}_t)$.
Step 4: Decoder Reconstruction
Initialize $Y_4 = \hat{Z}_4$. For $s = 4 \to 1$:
  Upsample $Y_s \to Y_{s-1}$.
  Fuse with skip S_{s-1} using ConvFuse.
Step 5: Output Layer
Reconstruct full-resolution Y_0.
Compute logits: $\hat{Y} = \text{Conv}_{1\times 1}(Y_0)$, apply sigmoid.
Step 6: Loss Optimization
Compute $\mathscr{L} = \lambda_{\text{dice}}\mathscr{L}_{\text{Dice}} + \lambda_{\text{ce}}\mathscr{L}_{\text{CE}}$.
Update weights with AdamW optimizer.

B. Experimental Setup

We evaluated the proposed SwinTextUNet on the QaTa-COV19 dataset [23] using the official split of 5,716 training, 1,429 validation, and 2,113 test images. All chest X-rays and lesion masks were resized to 224×224, intensity-normalized to $[0,1]$, and masks were binarized. Text annotations were standardized into concise diagnostic prompts and processed through the CLIP tokenizer. The model was implemented in PyTorch and trained end-to-end for 100 epochs using the AdamW optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-2}), cosine annealing with warmup, and a batch size of 8. A hybrid loss that combined dice and cross-entropy was used in the optimization process. Random flips, rotations, and intensity scaling were used for data augmentation.

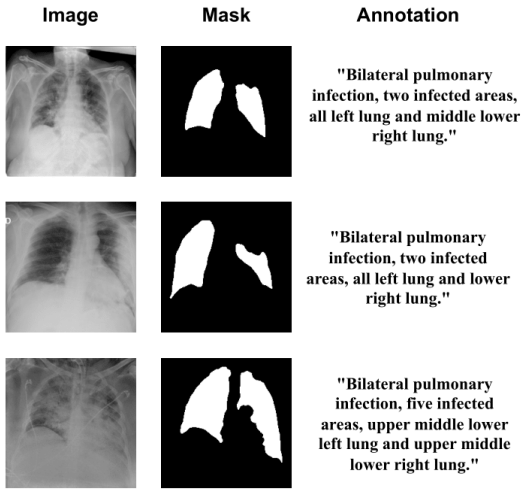

Fig. 5. Representative samples from the QaTa-COV19 dataset across training, validation, and test sets. Each sample shows (a) chest X-ray, (b) ground-truth lesion mask, and (c) textual annotation provided by the dataset authors [23].

C. Evaluation Metrics

We evaluate segmentation performance using the Dice Similarity Coefficient (Dice) and Intersection over Union (IoU), which quantify spatial overlap between predicted masks and ground truth. They are defined as:

$$Dice = \frac{2TP}{2TP+FP+FN}, \qquad IoU = \frac{TP}{TP+FP+FN}, \quad (13)$$

where TP, FP, and FN denote true positives, false positives, and false negatives. Dice emphasizes region overlap, while

IoU captures pixel-level agreement, making them well-suited for medical image segmentation.

D. Result analysis

This section provides a detailed evaluation of SwinTextUNet on the QaTa-COV19 dataset. The analysis covers segmentation accuracy, convergence behavior, architectural depth, ablation experiments, and comparisons with established baselines.

1) Stage-Depth Exploration: To investigate the effect of architectural depth, we experimented with 3-stage, 4-stage, and 5-stage variants of SwinTextUNet. The results, summarized in Table I, indicate that the 3-stage variant provides acceptable accuracy but fails to capture fine-grained structures, leading to reduced Dice and IoU scores. The 5-stage variant slightly improves accuracy but introduces considerable parameter overhead, resulting in diminishing returns. By contrast, the 4-stage architecture achieves the best trade-off, with a Dice score of 86.47% and IoU of 78.2%, while maintaining a balanced parameter count. This confirms that a four-stage encoder–decoder structure is the most suitable configuration for the segmentation task.

TABLE I
PERFORMANCE OF SWINTEXTUNET WITH DIFFERENT STAGE DEPTHS.

Variant	Dice (%)	IoU (%)	Params (M)
3-Stage	75.4	68.0	32.8
4-Stage (Ours)	**86.47**	**78.2**	57.6
5-Stage	87.1	78.8	105.4

2) Training Convergence: The convergence behavior of the 4-stage SwinTextUNet is illustrated in Figure 6. Both training and validation losses exhibit a smooth and monotonic decrease, with the validation loss closely following the training loss, suggesting minimal overfitting. Compared with 3-stage and 5-stage configurations, the 4-stage model demonstrates more stable convergence, which is consistent with its superior quantitative performance.

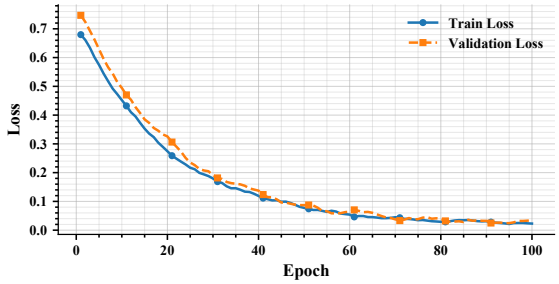

Fig. 6. Training and validation loss curves of SwinTextUNet.

3) Qualitative Results: Representative qualitative results are presented in Figure 7. Each triplet shows an original chest X-ray, its ground-truth lesion mask, and the corresponding prediction from SwinTextUNet. The model accurately captures lesion boundaries, including complex bilateral and multi-focal infections. Predictions remain robust across varying lesion morphologies and sizes. Compared with standard Swin-UNet outputs, SwinTextUNet produces sharper boundaries and fewer false positives, highlighting the value of text-guided cross-attention in enhancing spatial alignment.

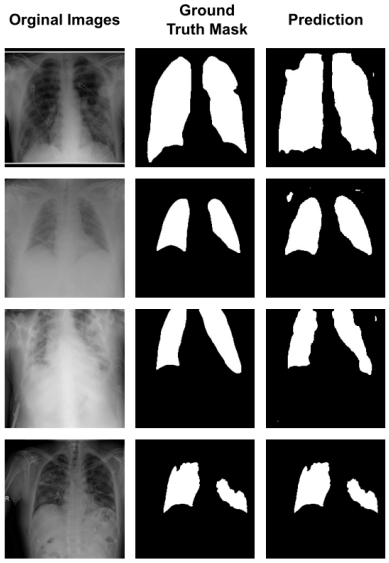

Fig. 7. Representative qualitative segmentation results. Each triplet shows (a) original CXR, (b) ground-truth mask, and (c) SwinTextUNet prediction.

4) Ablation Study: To examine the contribution of individual modules, we performed an ablation study by selectively disabling text guidance, cross-attention, and ConvFuse. Results in Table II show that removing text guidance leads to a substantial Dice drop of 7.3% and an IoU drop of 8.8%, underscoring the critical importance of semantic priors. Excluding ConvFuse decreases Dice by 3.1% and IoU by 7.0%, highlighting its role in effective multi-scale feature integration. Replacing cross-attention with simple concatenation causes a 1.8% Dice and 2.8% IoU reduction, demonstrating the necessity of explicit token-level alignment between modalities. The full SwinTextUNet consistently achieves the best results, validating the synergistic effect of all modules.

TABLE II
ABLATION STUDY OF SWINTEXTUNET COMPONENTS.

Model Variant	Dice (%)	IoU (%)
w/o Text Guidance	79.2	69.4
w/o ConvFuse	83.4	71.2
w/o Cross-Attention	84.7	75.4
Full SwinTextUNet	**86.47**	**78.2**

5) Comparison with Baseline Models: Finally, SwinTextUNet was compared against widely used segmentation models. As shown in Table III, traditional CNN-based approaches underperform due to limited contextual modeling. Transformer-based methods achieve stronger results, yet Swin-

TextUNet surpasses all baselines, reaching 86.47% Dice and 78.2% IoU.

TABLE III
THIS SECTION PROVIDES A DETAILED EVALUATION OF SWINTEXTUNET ON THE QATA-COV19 DATASET.

Model	Dice (%)	IoU (%)
U-Net [3]	79.02	69.46
Attention U-Net [24]	79.62	70.25
CLIP [2]	79.81	70.66
LViT [23]	83.66	75.71
SwinTextUNet (Ours)	**86.47**	**78.2**

E. Discussion

Overall, SwinTextUNet demonstrates consistent improvements over both CNN- and Transformer-based baselines. The 4-stage configuration emerges as the most effective depth, balancing performance and complexity. Ablation experiments emphasize the necessity of multimodal fusion components, while loss curve analysis confirms stable training dynamics. Overall, the results confirm the effectiveness of SwinTextUNet and its promise for clinical use where precision and interpretability are crucial.

V. CONCLUSION

We proposed SwinTextUNet, a multimodal segmentation framework that integrates CLIP-based textual embeddings into a Swin Transformer U-Net. Experiments on the QaTa-COV19 dataset showed consistent improvements over CNN- and Transformer-based baselines, with the four-stage variant offering the best trade-off between accuracy and complexity. This study has some limitations: experiments were limited to a single dataset and 2D images, and text annotations may not always be available in clinical settings. In future work, we aim to extend our model to multi-disease and 3D datasets, leverage domain-specific language models, and validate performance in real-world clinical workflows.

REFERENCES

[1] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*. Springer, 2018, pp. 3–11.

[5] A. Yeafi and L. Sarker, "Adtnet: Attention-guided u-net with dynamic cnn and transformers for skin cancer detection," in *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 2024, pp. 679–684.

[6] M. T. Jawad, A. Yeafi, and K. K. Halder, "Gsnet: a multi-class 3d attention-based hybrid glioma segmentation network," *Optics Express*, vol. 31, no. 24, pp. 40 881–40 906, 2023.

[7] A. Yeafi, M. Islam, and M. S. U. Yusuf, "A deep learning framework for 3d brain tumor segmentation and survival prediction," *Healthcare Analytics*, p. 100418, 2025.

[8] P. Goswami and A. A. Hossain, "Street object detection from synthesized and processed semantic image: A deep learning based study," *Human-Centric Intelligent Systems*, vol. 3, no. 4, pp. 487–507, 2023.

[9] P. Goswami, A. A. Hossain, and A. N. M. Sakib, "An end-to-end web-based system for rice leaf disease classification using deep learning," in *International Joint Conference on Advances in Computational Intelligence*. Singapore: Springer Nature Singapore, 2022, pp. 517–531.

[10] P. Goswami, A. A. Safi, A. N. M. Sakib, and T. Datta, "Corn leaf disease identification via transfer learning: A comprehensive web-based solution," in *International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology*. Singapore: Springer Nature Singapore, 2023, pp. 429–441.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[14] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[15] Q. Qi, L. Lin, R. Zhang, and C. Xue, "Medt: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 28 750–28 759, 2022.

[16] L. Sarker and A. Yeafi, "Ef-swinnet: A hybrid efficientnet-swin transformer model for skin cancer classification," in *2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET)*. IEEE, 2024, pp. 1–4.

[17] M. K. Islam, A. Biswas, and H. Hu, "Adaptive real-time gap detection system: A multi-algorithm approach integrating ultrasonic sensing and machine learning for robust structural analysis," in *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*. IEEE, 2025.

[18] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 016–15 027.

[19] S. Eslami, C. Meinel, and G. De Melo, "Pubmedclip: How much does clip benefit visual question answering in the medical domain?" in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1181–1193.

[20] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Medclip-sam: Bridging text and image towards universal medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2024, pp. 643–653.

[21] A. Yeafi, M. Islam, S. K. Mondal, K. I. H. Nashad, and M. S. U. Yusuf, "A semi-supervised approach for brain tumor classification using wasserstein generative adversarial network with gradient penalty," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2023, pp. 1–6.

[22] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2306–2310.

[23] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, vol. 43, no. 1, pp. 96–107, 2023.

[24] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.