HO CHI MINH CITY UNIVERSITY OF FOREIGN LANGUAGES
AND INFORMATION TECHNOLOGY (HUFLIT)
FACULTY OF INFORMATION TECHNOLOGY

GRADUATION PROJECT PROPOSAL

TransUNet for Automated Pancreas Segmentation from CT Scans

A Hybrid CNN-Transformer Deep Learning Approach

Student: Danh Hoang Hieu Nghi
Student ID: 23DH112270
Major: Information Technology
Supervisor: MSc. Vo Thi Hong Tuyet
Email: 23dh112270@st.huflit.edu.vn

Ho Chi Minh City, January 2026

Contents

1 Introduction

1.1 Background and Motivation

Pancreatic cancer remains one of the most lethal malignancies worldwide, with a 5-year survival rate of approximately 10%. Early detection and accurate diagnosis play a decisive role in treatment outcomes, and precise segmentation of the pancreas from CT (Computed Tomography) images is the critical first step.

However, automated pancreas segmentation presents significant challenges:

- **Small organ size:** The pancreas occupies less than 1% of the total abdominal scan volume
- **High shape variability:** Significant anatomical differences between patients
- **Low contrast:** Poor differentiation from surrounding soft tissues
- **Variable position:** Location changes based on patient positioning and physiological state

Recently, the **TransUNet** architecture [1] has demonstrated superior performance in medical image segmentation by combining CNNs (Convolutional Neural Networks) with Vision Transformers. This hybrid approach addresses the limitations of pure CNN methods in capturing long-range dependencies while maintaining the ability to extract fine-grained local features essential for accurate boundary delineation.

1.2 Research Questions

This project aims to address the following research questions:

RQ1: How does the hybrid CNN-Transformer architecture (TransUNet) compare to traditional CNN-based methods (U-Net, Attention U-Net) in terms of pancreas segmentation accuracy on the MSD Task07 dataset?

RQ2: What is the trade-off between model complexity (parameters, computational cost) and segmentation performance for different TransUNet configurations?

RQ3: Can memory-efficient training strategies enable effective TransUNet training on consumer-grade GPUs (4-8GB VRAM) without significant performance degradation?

RQ4: How do different preprocessing techniques and data augmentation strategies affect the model's ability to handle the high variability in pancreas shape and position?

1.3 Hypotheses

H1: TransUNet will achieve at least 5% improvement in Dice score compared to baseline U-Net due to its ability to capture global context through self-attention mechanisms.

H2: The TransUNet-Base configuration will provide the optimal balance between performance and computational efficiency for clinical deployment scenarios.

H3: Gradient checkpointing and mixed-precision training will reduce memory requirements by at least 40% while maintaining model accuracy within 1% of full-precision training.

1.4 Objectives

Primary Objective:

To implement and evaluate the TransUNet architecture for automated pancreas segmentation from CT scans, achieving competitive performance with state-of-the-art methods.

Specific Objectives:

1. Research and implement the complete TransUNet architecture
2. Develop an optimized data preprocessing pipeline for pancreas CT images
3. Design memory-efficient training strategies for consumer-grade GPUs
4. Conduct quantitative and qualitative evaluation of model performance
5. Provide open-source code and comprehensive documentation

1.5 Scope

- **Dataset:** Medical Segmentation Decathlon Task07 Pancreas (420 volumes)
- **Approach:** 2D slice-based training
- **Framework:** PyTorch 2.0+, MONAI 1.3+
- **Minimum Hardware:** NVIDIA GPU with 4GB+ VRAM

1.6 Significance and Contributions

This research makes the following contributions:

1. **Theoretical Contribution:** Systematic evaluation of CNN-Transformer hybrid architectures for the challenging task of pancreas segmentation, providing insights into the effectiveness of self-attention mechanisms for small organ segmentation with high shape variability.

2. **Practical Contribution:** Development of memory-efficient training strategies (gradient checkpointing, mixed-precision training) that enable state-of-the-art model training on consumer-grade hardware, democratizing access to advanced medical imaging AI.

3. **Methodological Contribution:** Comprehensive preprocessing pipeline specifically optimized for pancreas CT images, including adaptive HU windowing and intelligent slice selection strategies.

4. **Community Contribution:** Open-source implementation with detailed documentation, pre-trained models, and tutorial notebooks to facilitate reproducibility and further research in medical image segmentation.

2 Literature Review

2.1 Deep Learning for Medical Image Segmentation

U-Net [2] established the encoder-decoder architecture with skip connections as the standard for medical image segmentation. Subsequent improvements include:

- **Attention U-Net** [4]: Adding attention mechanisms to skip connections
- **U-Net++**: Dense skip connections
- **ResUNet**: Integrating residual connections

2.2 Vision Transformer

Vision Transformer (ViT) [3] applies the self-attention mechanism to computer vision, enabling models to learn long-range relationships between image regions. The self-attention formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

2.3 Hybrid CNN-Transformer Architectures

TransUNet [1] combines the advantages of both approaches:

- **CNN:** Efficient local feature extraction
- **Transformer:** Global context modeling
- **U-Net Decoder:** Resolution recovery with skip connections

Other notable hybrid architectures include UNETR [6] which uses a pure Transformer encoder, and Swin-UNet which employs shifted window attention for computational efficiency.

2.4 Comparison of State-of-the-Art Methods

Table 1: Comparison of segmentation methods for pancreas (reported on various datasets)

Method	DSC	HD95	Year	Architecture
U-Net [2]	0.71-0.76	18-25mm	2015	CNN
Attention U-Net [4]	0.74-0.79	12-18mm	2018	CNN+Attention
nnU-Net	0.79-0.84	8-15mm	2021	Self-configuring CNN
TransUNet [1]	0.80-0.85	6-12mm	2021	CNN+Transformer
UNETR [6]	0.78-0.83	8-14mm	2022	Pure Transformer
Swin-UNet	0.79-0.84	7-13mm	2022	Swin Transformer

2.5 Research Gaps

Despite significant progress, several gaps remain in the literature:

1. **Limited accessibility:** Most state-of-the-art methods require high-end GPUs (16GB+ VRAM) for training, limiting adoption in resource-constrained settings such as hospitals in developing countries.
2. **Reproducibility issues:** Many published works lack complete implementation details or open-source code, making fair comparison difficult.
3. **2D vs 3D trade-offs:** While 3D methods capture volumetric context, they are computationally expensive. The optimal balance between 2D efficiency and 3D context for pancreas segmentation remains underexplored.

4. **Preprocessing standardization:** Different studies use varying preprocessing pipelines, making cross-study comparison challenging.

5. **Clinical deployment considerations:** Few studies address practical deployment aspects such as inference speed, memory footprint, and integration with clinical workflows.

This project addresses these gaps by providing an accessible, well-documented implementation with memory-efficient training strategies and comprehensive benchmarking.

3 Methodology

3.1 TransUNet Architecture

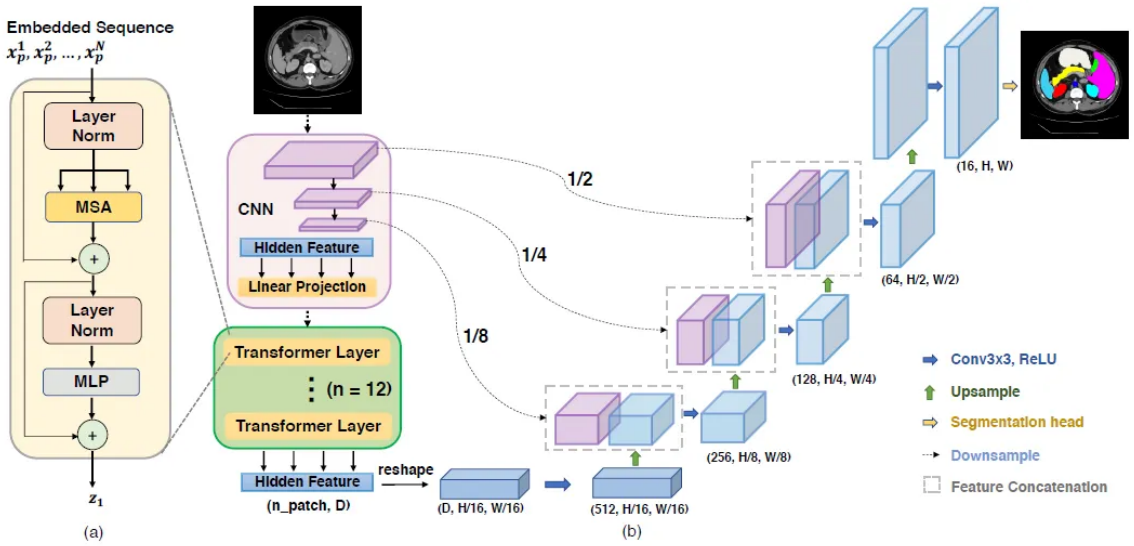

Figure 1: TransUNet architecture overview. The model combines a CNN encoder (ResNet-50) for local feature extraction with a Vision Transformer for global context modeling, followed by a cascaded upsampler decoder with skip connections.

Main Components:

1. **CNN Encoder:** 4 ResNet-style stages, output channels $\{64, 128, 256, 512\}$

2. **Vision Transformer:** 12 layers, hidden dim 768, 12 attention heads

3. **CNN Decoder:** 4 upsampling stages with skip connections

3.2 Loss Function

Using hybrid loss to address class imbalance:

$$\mathcal{L} = 0.5 \cdot \mathcal{L}_{\text{Dice}} + 0.5 \cdot \mathcal{L}_{\text{CE}} \tag{2}$$

3.3 Preprocessing Pipeline

1. Load NIfTI and standardize orientation (RAS)
2. Resample to isotropic spacing ($1.0 \times 1.0 \times 1.0$ mm)
3. HU windowing: clip [-175, 250], normalize to [0, 1]
4. Foreground cropping
5. Extract 2D axial slices
6. Resize to 224×224

3.4 Experimental Design

3.4.1 Training Configuration

Table 2: Hyperparameter configuration

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	1×10^{-4} (with cosine annealing)
Weight Decay	1×10^{-4}
Batch Size	8-16 (depending on GPU memory)
Epochs	150 (with early stopping, patience=20)
Warmup Epochs	10

3.4.2 Data Augmentation

To improve model generalization and address limited training data:

- Random horizontal/vertical flipping (p=0.5)
- Random rotation ($\pm 15°$)
- Random scaling (0.9-1.1)

- Elastic deformation ($\alpha = 100$, $\sigma = 10$)
- Intensity augmentation: brightness (± 0.1), contrast (± 0.1)
- Random Gaussian noise ($\sigma = 0.01$)

3.4.3 Memory-Efficient Training Strategies

1. **Gradient Checkpointing:** Trade computation for memory by recomputing intermediate activations during backward pass, reducing memory by $\sim$40%.
2. **Mixed-Precision Training (FP16):** Using NVIDIA Automatic Mixed Precision (AMP) to reduce memory footprint and accelerate training.
3. **Gradient Accumulation:** Simulate larger batch sizes by accumulating gradients over multiple forward passes.

3.4.4 Cross-Validation Strategy

To ensure robust performance estimation:

- 5-fold cross-validation on the training set
- Stratified splitting to maintain pancreas-to-background ratio
- Final model trained on full training set, evaluated on held-out test set
- Statistical significance testing using paired t-test ($\alpha = 0.05$)

3.4.5 Ablation Studies

The following ablation experiments will be conducted:

1. Impact of Transformer depth (6, 12, 24 layers)
2. Effect of patch size (8, 16, 32)
3. Contribution of skip connections
4. Loss function comparison (Dice, CE, Dice+CE, Focal)
5. Preprocessing variations (HU window ranges, normalization methods)

Attribute	Value
Source	Memorial Sloan Kettering Cancer Center
Volume Count	420 CT volumes
Labels	Background (0), Pancreas (1), Tumor (2)
Format	NIfTI (.nii.gz)
Size	∼11.4 GB

Table 3: MSD Task07 Pancreas dataset information

4 Dataset

Medical Segmentation Decathlon (MSD) Task07 Pancreas:

Data Split:

- Training: 80% (336 volumes, ∼8,000 slices)
- Validation: 10% (42 volumes)
- Test: 10% (42 volumes)

5 Evaluation Metrics

5.1 Dice Similarity Coefficient (DSC)

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|} \tag{3}$$

Measures overlap between prediction and ground truth (0-1).

5.2 Hausdorff Distance (HD95)

$$\text{HD}(P, G) = \max\left(\sup_{p \in P} \inf_{g \in G} d(p, g), \sup_{g \in G} \inf_{p \in P} d(g, p)\right) \tag{4}$$

Measures boundary distance (mm), HD95 is the 95th percentile.

5.3 Intersection over Union (IoU)

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{5}$$

6 Expected Results

Based on the original TransUNet paper and related works on pancreas segmentation, we expect the following performance metrics on the test set (Table 6).

Model	DSC	HD95 (mm)	Params
U-Net (baseline)	0.70-0.75	15-20	31M
TransUNet-Small	0.75-0.78	12-16	17M
TransUNet-Base	0.78-0.83	8-12	105M
TransUNet-Large	0.82-0.86	5-10	300M

Table 4: Expected results on test set

The TransUNet-Base model is expected to achieve the best trade-off between accuracy and computational cost, making it suitable for potential clinical deployment.

7 Timeline

The project will be conducted over a 12-week period. The detailed timeline with tasks and deliverables is presented in Table 7.

Week	Task	Deliverable
1-2	Literature review	Survey document
3-4	Dataset preparation, preprocessing	Data pipeline
5-6	TransUNet implementation	Model code
7-8	Training and fine-tuning	Trained models
9-10	Evaluation and ablation study	Results analysis
11-12	Report writing, documentation	Final report

Table 5: Project timeline (12 weeks)

8 Deliverables

1. **Source Code:** GitHub repository with complete implementation
2. **Trained Models:** Checkpoints for Small/Base/Large variants
3. **Documentation:** README, API docs, tutorial notebooks
4. **Report:** Complete graduation project report

5. **Demo:** Jupyter notebooks and/or web demo

GitHub Repository:
https://github.com/ihatesea69/TransUNet-Pancreas-Segmentation

9 Limitations and Ethical Considerations

9.1 Technical Limitations

1. **2D Approach Limitation:** The 2D slice-based approach may miss important volumetric context that 3D methods can capture. This trade-off is made to enable training on consumer-grade hardware.

2. **Class Imbalance:** The pancreas occupies <1% of abdominal CT volume. While addressed through loss functions and sampling strategies, severe imbalance may still affect performance on boundary regions.

3. **Domain Shift:** The model trained on MSD Task07 may not generalize well to CT scans from different scanners, protocols, or patient populations without fine-tuning.

4. **Tumor Segmentation:** This project focuses primarily on pancreas parenchyma segmentation. Tumor segmentation (label 2) is secondary and may require additional specialized techniques.

9.2 Risk Mitigation

- Extensive validation on held-out test set
- Cross-validation to assess model stability
- Uncertainty estimation through Monte Carlo dropout
- Clear documentation of model limitations for end-users

9.3 Ethical Considerations

1. **Data Privacy:** The MSD dataset is publicly available and anonymized. No personally identifiable information (PII) is used or stored.

2. **Clinical Use Disclaimer:** This is a research project and the developed model is NOT intended for direct clinical diagnosis without proper validation by qualified medical professionals and regulatory approval.

3. **Bias and Fairness:** The MSD dataset may not represent global patient diversity. Model performance should be validated on diverse populations before any clinical application.

4. **Reproducibility:** All code, trained models, and documentation will be made publicly available to ensure reproducibility and facilitate peer review.

10 Conclusion

This proposal outlines a comprehensive research plan to implement and evaluate TransUNet for automated pancreas segmentation from CT scans. The project addresses key research questions regarding the effectiveness of hybrid CNN-Transformer architectures, computational efficiency trade-offs, and practical deployment considerations.

The expected contributions include: (1) a systematic benchmark of TransUNet variants on the MSD Task07 dataset, (2) memory-efficient training strategies enabling consumer-grade GPU training, (3) an optimized preprocessing pipeline for pancreas CT images, and (4) open-source implementation with comprehensive documentation.

Successful completion of this project will advance the state of medical image segmentation research while providing accessible tools for the broader research community.

References

[1] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv:2102.04306*, 2021.
https://arxiv.org/abs/2102.04306

[2] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.
https://arxiv.org/abs/1505.04597

[3] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
https://arxiv.org/abs/2010.11929

[4] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," *MIDL*, 2018.
https://arxiv.org/abs/1804.03999

[5] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv:1902.09063*, 2019.
https://arxiv.org/abs/1902.09063

[6] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," *WACV*, 2022.
https://arxiv.org/abs/2103.10504