UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Semantic Segmentation with U-Net and Its Variants

Vision and Cognitive Systems - Project Presentation

Oday Najad Wageesha Widuranga Tommaso Di Tullio

February 9, 2025

Table of Contents

Semantic Segmentation is a fundamental task in computer vision that involves classifying each pixel of an image into a predefined category. It is widely used in applications such as medical imaging, autonomous driving, and remote sensing.

Objective of this Project

The goal of this project is to explore and compare the performance of three different deep learning architectures—**U-Net, U-Net++, and Attention U-Net**—for semantic segmentation. The evaluation focuses on accuracy, generalization, and computational efficiency.

- Implement and train **U-Net**, **U-Net++**, and **Attention U-Net**.
- Analyze and compare their performance using standard segmentation metrics.
- Assess their suitability for real-world applications.

Several deep learning architectures have been proposed for semantic segmentation:

- **U-Net** [Ronneberger et al., 2015] was initially introduced for biomedical image segmentation and has been widely used in medical imaging and satellite image analysis.
- **U-Net++** [Zhou et al., 2018] improved upon U-Net by adding dense skip pathways and deep supervision, leading to enhanced segmentation performance.
- **Attention U-Net** [Oktay et al., 2018] introduced attention mechanisms to focus on relevant image regions, improving segmentation accuracy in medical imaging and remote sensing applications.

These architectures serve as the foundation for our project, where we compare their effectiveness on the given dataset.

Dataset (Binary Segmentation): Colorectal Cancer WSI Dataset

The **Colorectal Cancer WSI Dataset** was employed to evaluate the models on a binary segmentation task. This dataset comprises 2,198 images, split into:

- **Training Set**: 1,535 images with corresponding binary masks.
- **Validation Set**: 663 images for performance evaluation.

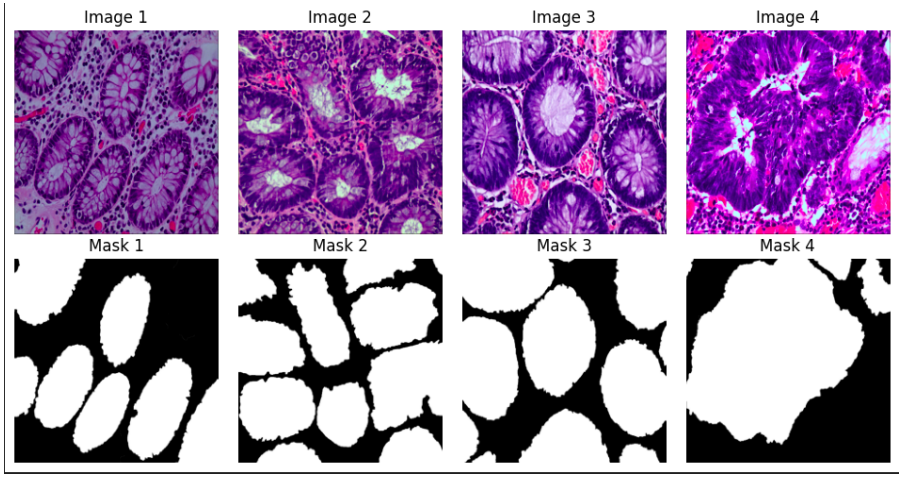

Figure: Binary Segmentation sample.

The images represent different tumor types with corresponding segmentation masks, enabling the models to differentiate between tumor and non-tumor regions.

Dataset (Multi-class Segmentation): Cityscapes Dataset

The dataset used in this project is the **Cityscapes dataset**, a widely recognized benchmark for urban scene understanding.

- Composed of high-resolution **street scene images** from **50 different cities**.
- Each image is accompanied by **pixel-wise labeled segmentation masks**.
- Classes include **roads, buildings, pedestrians, vehicles, trees, etc.**

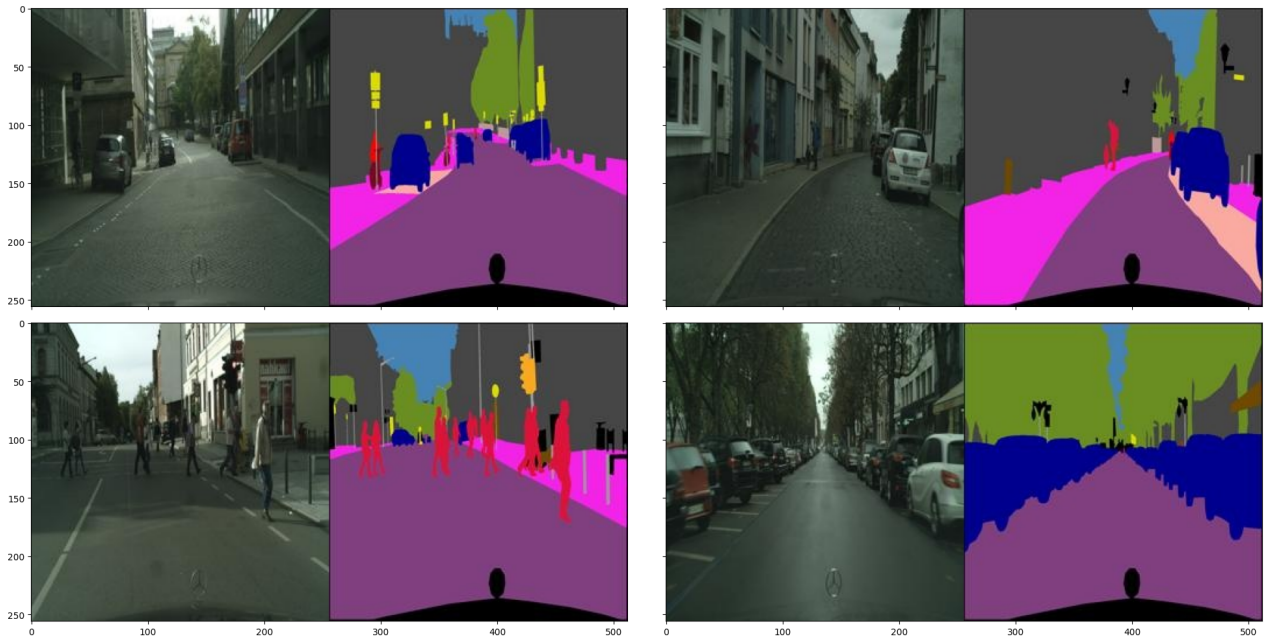

Figure: Example batch of input images and their corresponding segmentation masks.

Dataset Splitting

The dataset was divided into two sets:

- **Training Set**: 80% of the dataset for model learning.
- **Validation Set**: 20% of the dataset for evaluating generalization.

Why this split?
- Ensures that the model **learns robust features** while retaining **sufficient validation data** for unbiased evaluation.

Data Preprocessing

Several preprocessing steps were applied to standardize the dataset:

- **Resizing**: All images were resized to **128×128** for computational efficiency.
- **Normalization**: Pixel values were normalized to **[0,1]** to accelerate model convergence.
- **Categorical Mapping**: Masks were originally stored as RGB images, so we converted them into **integer class indices** for efficient processing.
- **One-Hot Encoding**: Necessary for multi-class segmentation to allow per-pixel classification.
- **Data Augmentation**: Techniques like **random cropping, flipping, and brightness adjustment** were applied to improve generalization.

Why these steps?

- **Normalization** improves training stability. - **Categorical mapping** reduces memory overhead. - **Augmentation** increases robustness against overfitting.

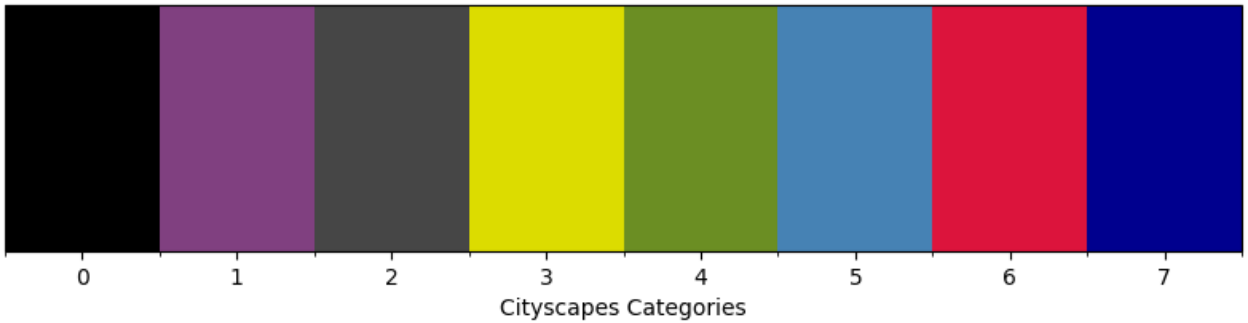

Figure: Visualization of different semantic segmentation classes.

U-Net Architecture

U-Net is a fully convolutional neural network originally designed for biomedical image segmentation. It consists of:

- **Encoder Path:** Captures spatial features using convolutional and pooling layers.
- **Decoder Path:** Uses transposed convolutions to recover spatial resolution.
- **Skip Connections:** Bridge corresponding encoder-decoder layers to retain fine-grained details.

Advantage: Effective for small dataset sizes due to strong feature propagation through skip connections.

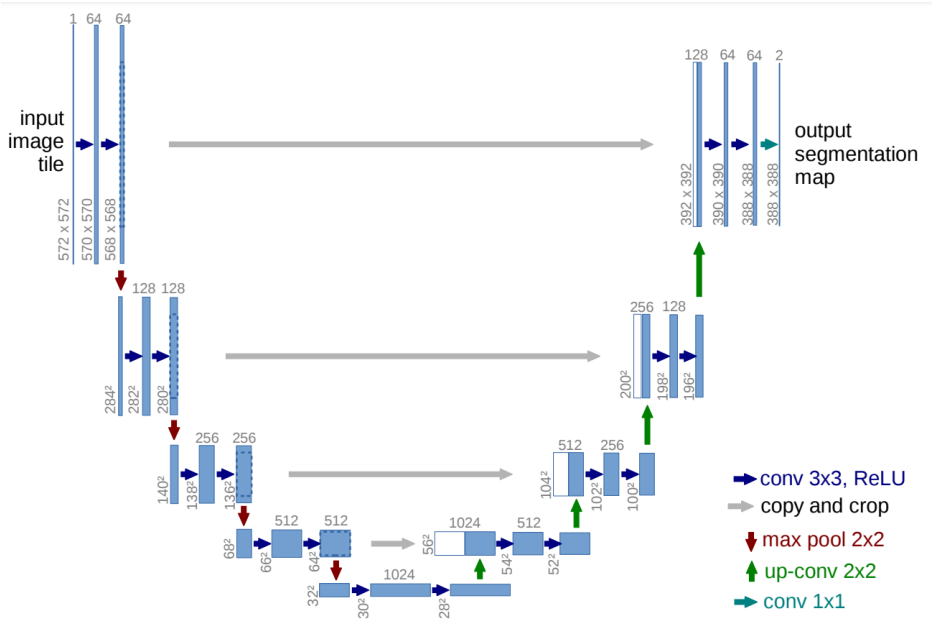

Figure: Example batch of input images and their corresponding segmentation masks.

Nested U-Net (U-Net++) Architecture

U-Net++ is an enhanced version of U-Net with dense connections that refine feature learning. It introduces:

- **Nested Skip Pathways:** Progressive feature fusion across multiple intermediate layers.
- **Deep Supervision:** Enhances gradient flow and stabilizes training.

Advantage: Improves segmentation accuracy by reducing semantic gaps between encoder and decoder features.

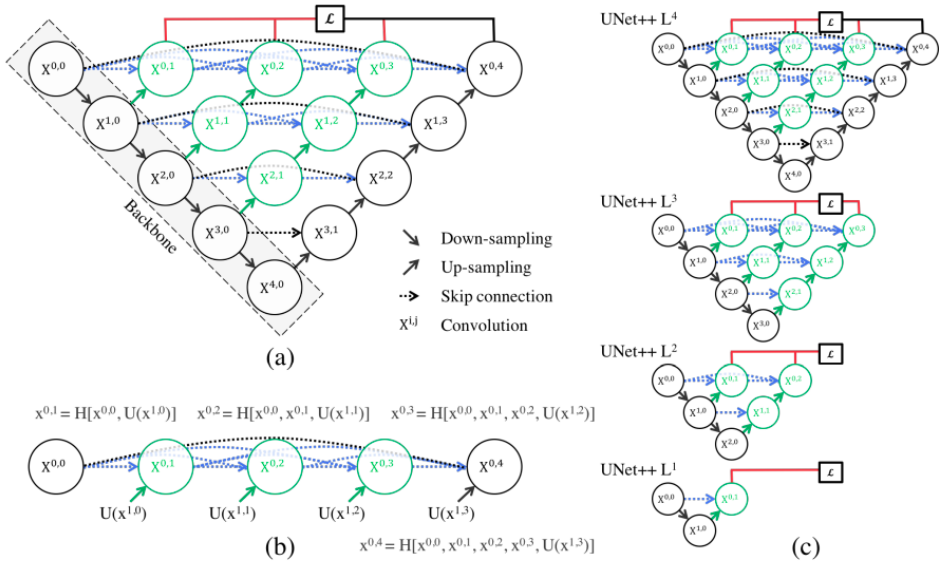

Figure: U-Net++ Architecture with nested skip pathways and deep supervision.

Attention U-Net Architecture

Attention U-Net integrates an attention mechanism to dynamically refine feature maps. It includes:

- **Attention Gates:** Suppresses irrelevant background features while enhancing crucial details.
- **Context-Aware Feature Selection:** Helps the model focus on important regions in the image.

Advantage: Increases segmentation precision, especially for complex structures with low contrast.

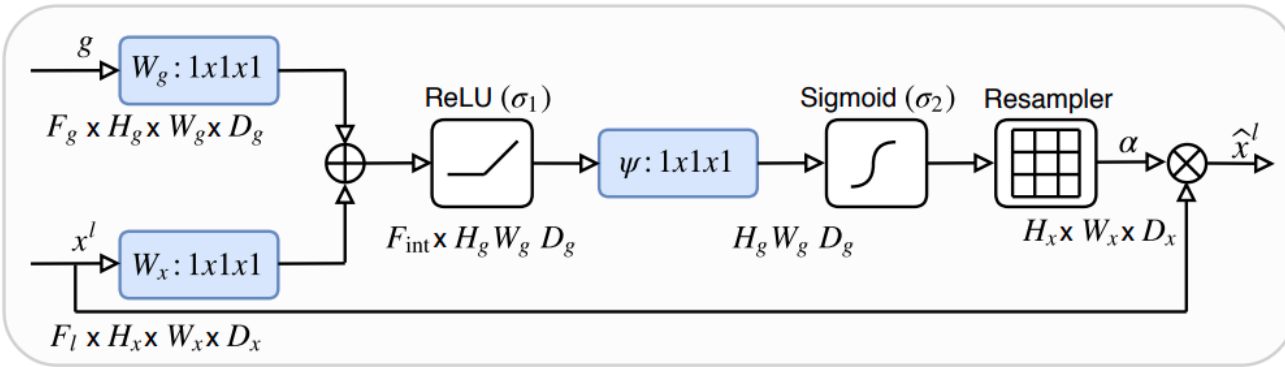

Figure: Attention U-Net Architecture with attention gates to refine feature selection.

Training Process

The models were trained using a combination of optimization techniques, loss functions, and data augmentation strategies to enhance generalization and performance.

Optimizer: Adam

- Adaptive moment estimation (Adam) optimizer was used for training.
- **Advantage:** Combines the benefits of momentum-based and adaptive learning rate methods.
- Helps the model converge faster while avoiding vanishing or exploding gradients.

Loss Function: Dice + Cross-Entropy Loss

- **Cross-Entropy Loss:** Measures per-pixel classification loss for segmentation.
- **Dice Loss:** Measures the overlap between predicted and true masks.
- **Advantage:** Balances between pixel-wise accuracy and segmentation shape consistency.

Data Augmentation

- Applied transformations: **horizontal flipping, rotation, brightness scaling**.
- **Purpose:** Increases dataset variability and prevents overfitting.

Performance on Binary Segmentation: U-Net

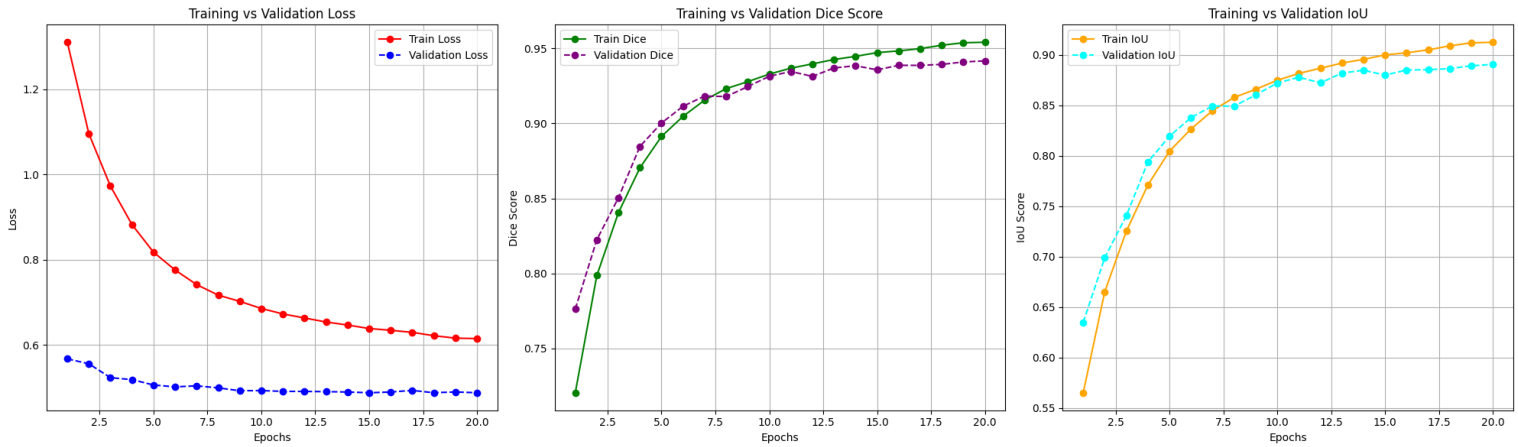

Figure: Training and validation performance of U-Net.

The plot indicates a steady decrease in loss with improved Dice and IoU scores, achieving a **Dice score of 0.942** and **IoU of 0.890**. The model shows strong generalization with minimal overfitting.

U-Net Inference Results

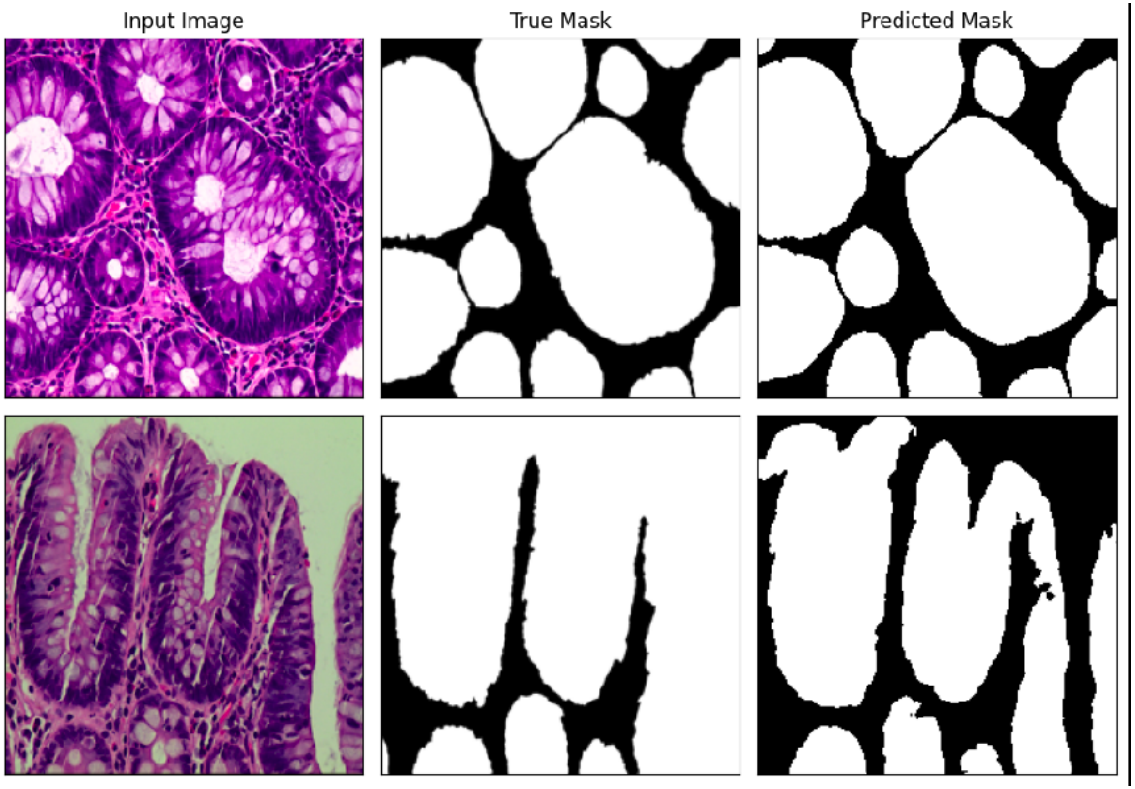

Figure: Inference results using U-Net.

The predictions align well with the ground truth masks, capturing fine details of tumor boundaries accurately.

Nested U-Net Performance

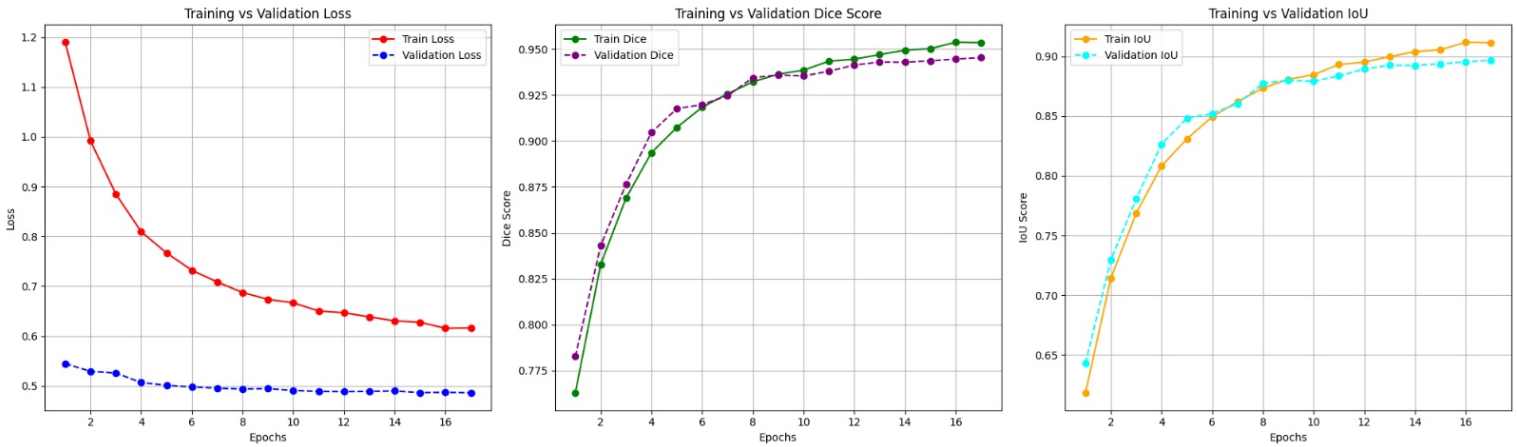

Figure: Training and validation performance of Nested U-Net (U-Net++).

Nested U-Net achieved a **Dice score of 0.945** and **IoU of 0.897**. The performance is comparable to U-Net, with slightly faster convergence due to enhanced feature propagation.

Nested U-Net Inference Results

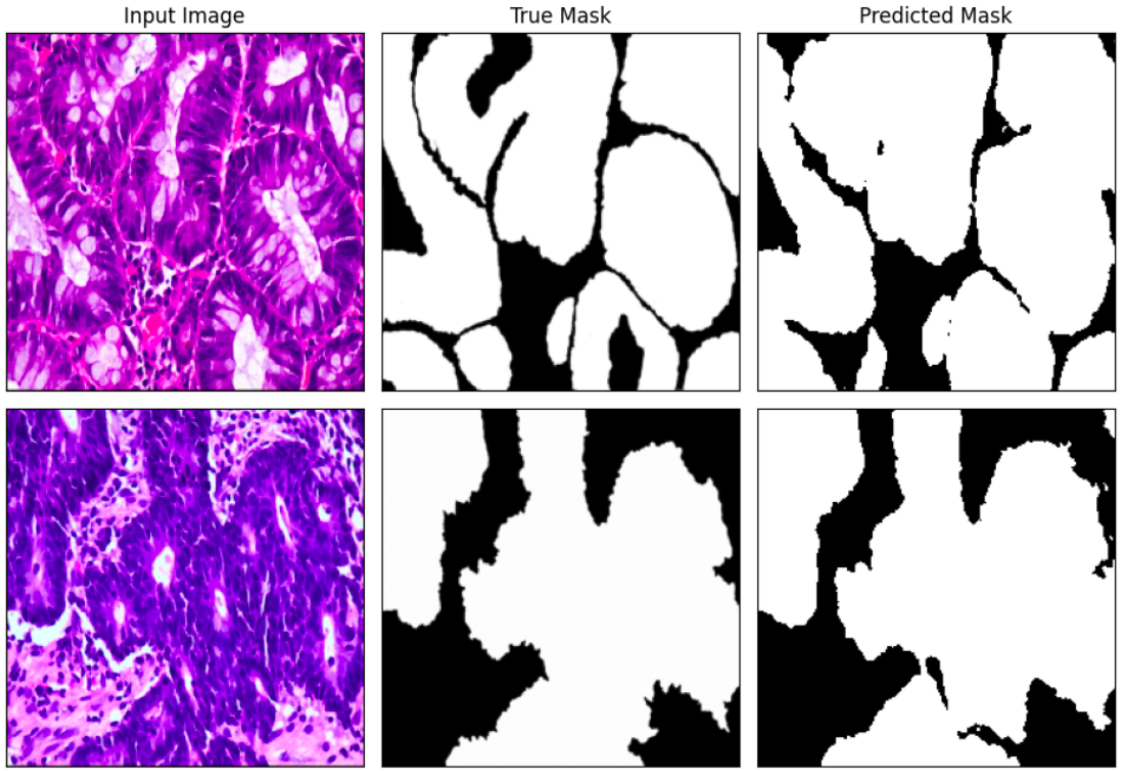

Figure: Inference results using Nested U-Net (U-Net++).

The model effectively segments tumor regions, demonstrating robust performance in complex tissue structures.

Attention U-Net Performance

Figure: Training and validation performance of Attention U-Net.

Attention U-Net achieved a **Dice score of 0.931** and **IoU of 0.871**. The attention mechanism improved focus on relevant tumor regions, although slight performance degradation was noted compared to U-Net and U-Net++.

Attention U-Net Inference Results

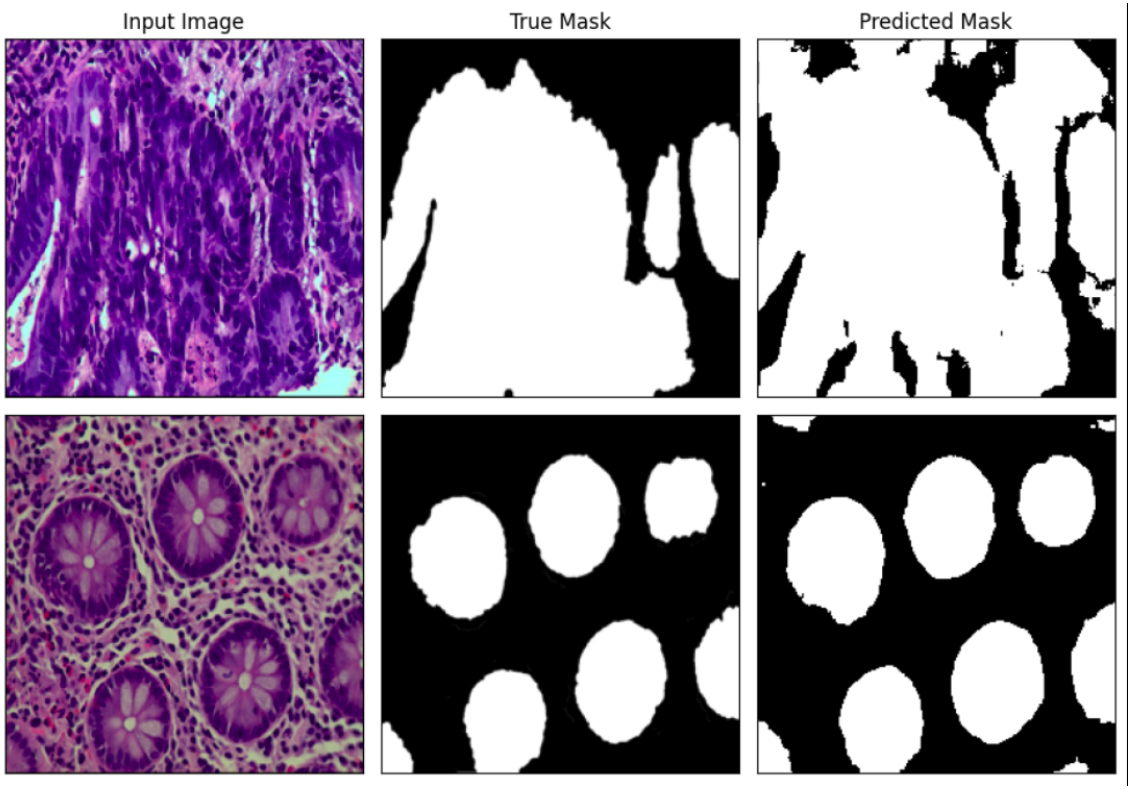

Figure: Inference results using Attention U-Net.

The model demonstrates good segmentation quality but shows minor misclassification in boundary areas compared to U-Net and U-Net++.

Performance Comparison

Table: Binary Segmentation Performance Comparison

Model	Dice Score	IoU Score
U-Net	0.942	0.890
Nested U-Net (U-Net++)	0.945	0.897
Attention U-Net	0.931	0.871

Multi-Class: U-Net Performance

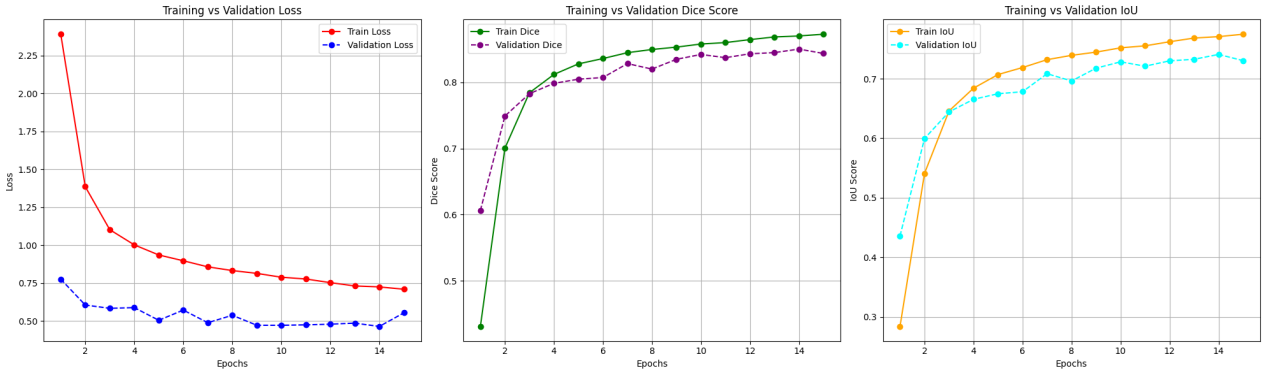

Figure: Training performance of U-Net over the epochs.

Analysis:

- U-Net achieved stable convergence after multiple epochs.

Multi-Class: U-Net Performance

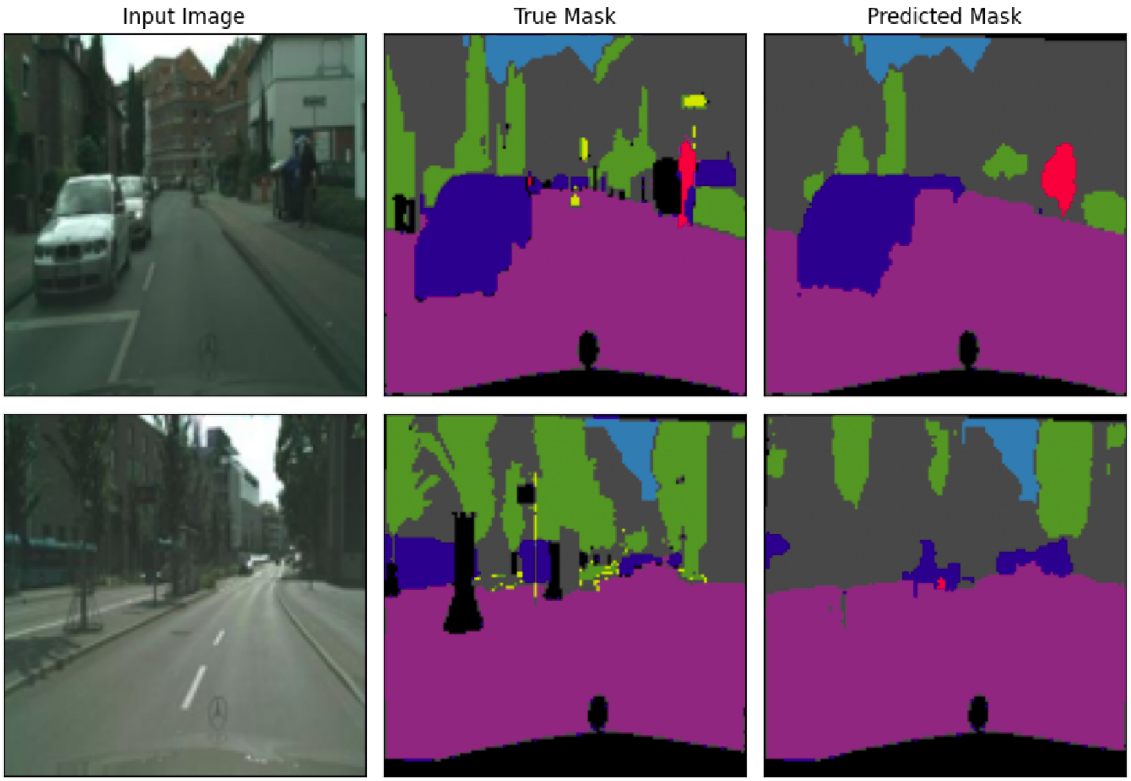

Figure: Inference results using U-Net.

Analysis:

- The segmentation masks produced reasonable outlines, but some fine details were lost.
- **Limitation:** Struggles with intricate regions due to limited feature refinement.

Multi-Class: Nested U-Net (U-Net++) Performance

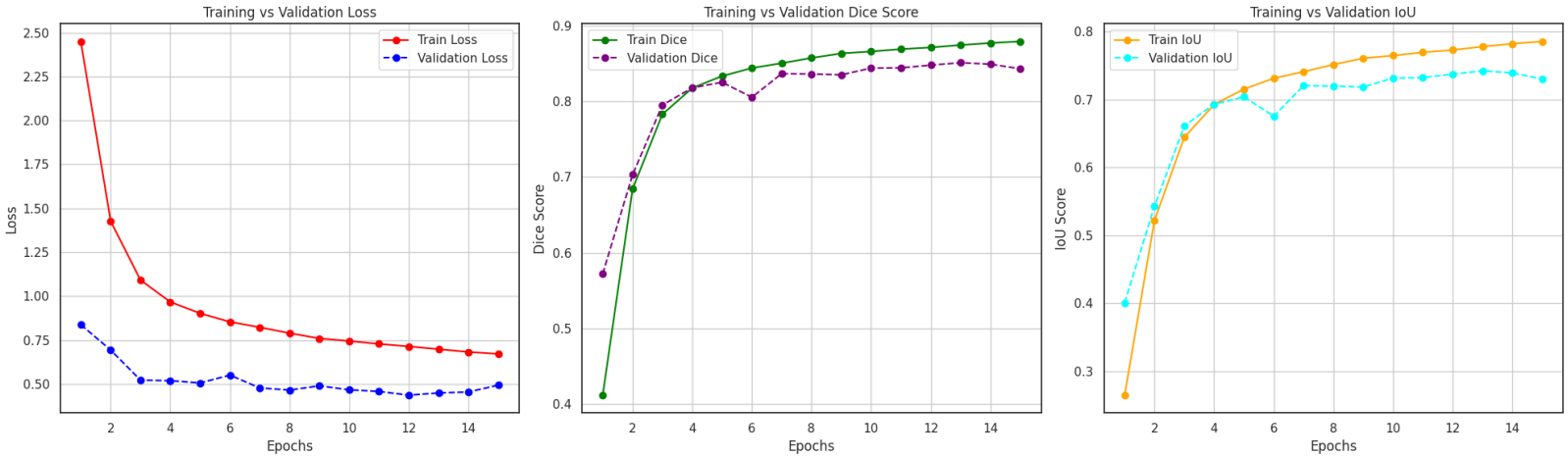

Figure: Training performance of Nested U-Net (U-Net++).

Analysis:

- U-Net++ demonstrated faster convergence and lower loss values.

Multi-Class: Nested U-Net (U-Net++) Performance

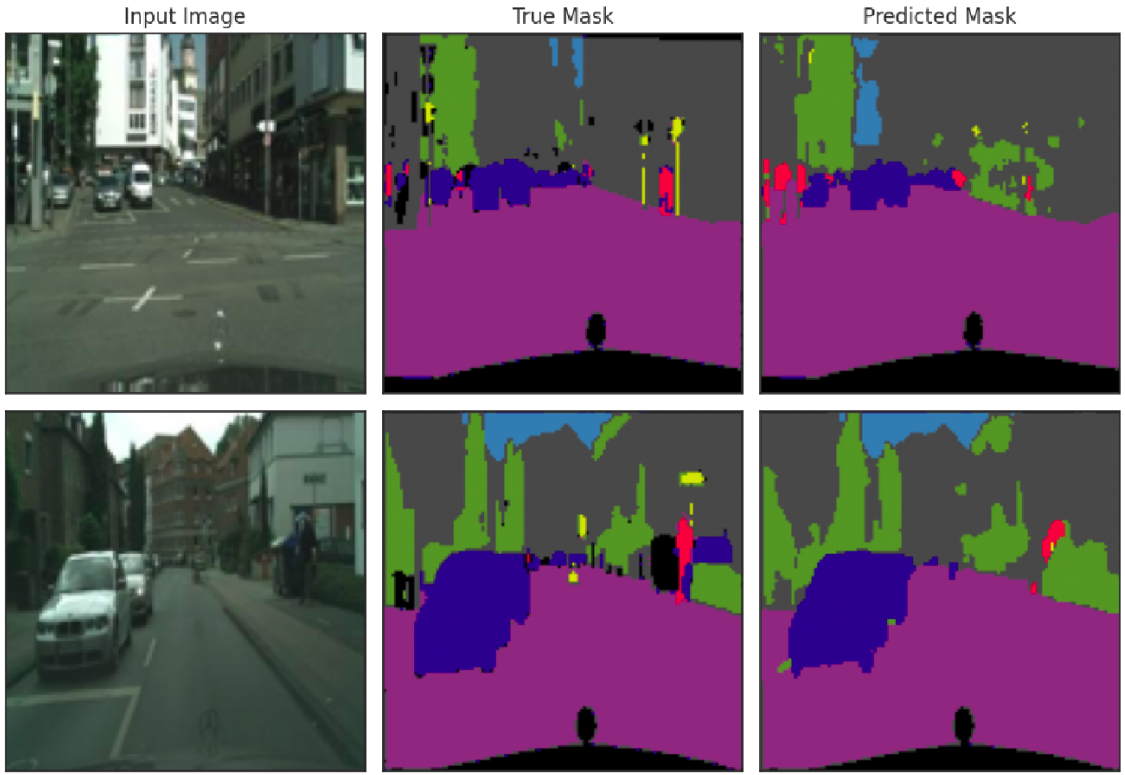

Figure: Inference results using Nested U-Net (U-Net++).

Analysis:

- More refined segmentation masks compared to standard U-Net.
- **Strength:** Reduces semantic gaps with densely connected pathways.

Multi-Class: Attention U-Net Performance

Figure: Training performance of Attention U-Net.

Analysis:

- Notably, the validation Dice Score stabilizes at a higher value compared to the other models.

Multi-Class: Attention U-Net Performance

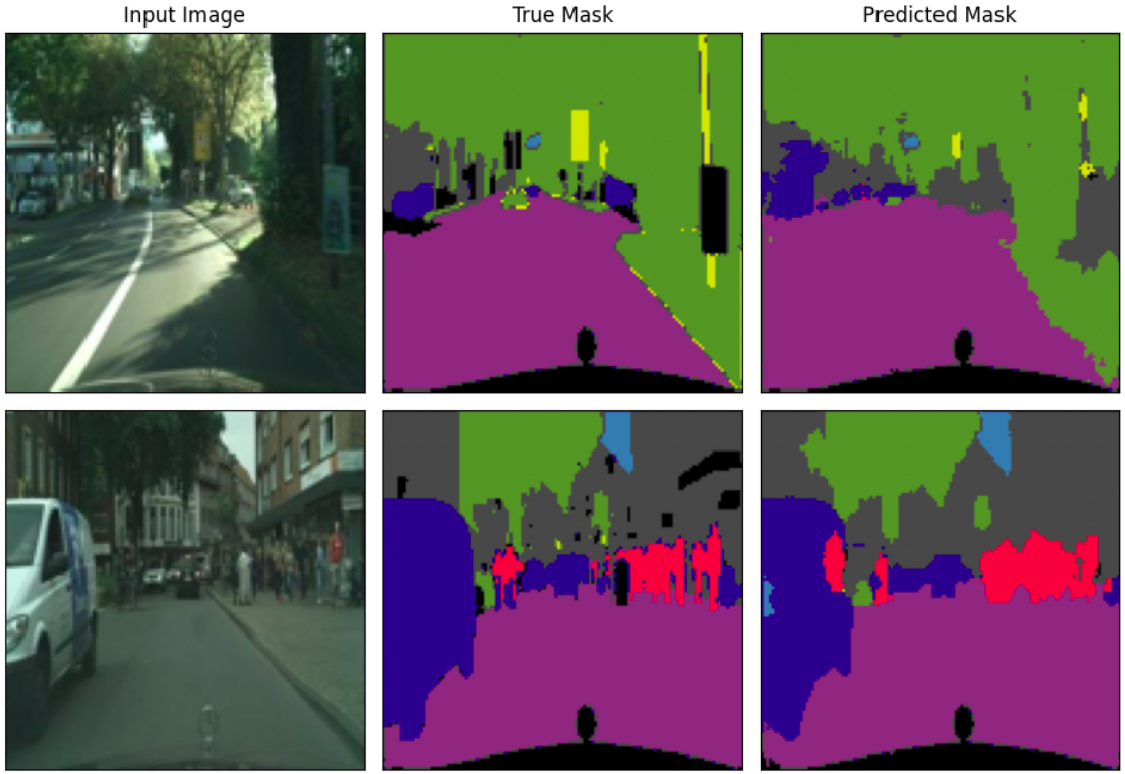

Figure: Inference results using Attention U-Net.

Analysis:

- Attention U-Net further improved segmentation precision.
- Attention mechanisms helped suppress irrelevant background features.
- **Advantage:** Performs well on complex structures with low contrast.

While our models achieved competitive results, state-of-the-art segmentation architectures such as **DeepLabV3+** [Chen et al., 2018] and **PSPNet** [Zhao et al., 2017] have demonstrated superior performance due to advanced network architectures and feature aggregation techniques.

Table: Comparison of our models with state-of-the-art segmentation architectures.

Model	Dice Score	IoU Score
U-Net	0.851	0.742
U-Net++	0.857	0.750
Attention U-Net	0.864	0.758
DeepLabV3+ [Chen et al., 2018]	0.910	0.805
PSPNet [Zhao et al., 2017]	0.902	0.798

Why Do DeepLabV3+ and PSPNet Outperform Our Models?

- **Atrous Spatial Pyramid Pooling (ASPP)**: DeepLabV3+ uses ASPP to extract features at multiple scales, capturing both fine and coarse details.
- **Global Context Aggregation**: PSPNet employs a **pyramid pooling module** that enhances the model's ability to understand large-scale structures in the image.
- **Stronger Backbone Networks**: Both models use **ResNet-101** or **Xception** backbones, which are deeper and more powerful than standard U-Net encoders.
- **Better Generalization**: DeepLabV3+ and PSPNet are pre-trained on large-scale datasets like **COCO** and **PASCAL VOC**, allowing them to generalize better to unseen images.
- **Post-processing Techniques**: These models integrate techniques such as **Conditional Random Fields (CRFs)** to refine segmentation boundaries.

Summary of Findings

- This project explored **semantic segmentation** using **U-Net**, **Nested U-Net (U-Net++)**, and **Attention U-Net**.
- We demonstrated that incorporating **dense skip connections** (U-Net++) and **attention mechanisms** (Attention U-Net) enhances segmentation accuracy.
- Extensive **experiments** showed that **Attention U-Net** achieved the highest **Dice** and **IoU scores**, outperforming standard U-Net.
- In the **binary segmentation task**, **U-Net** and **U-Net++** slightly outperformed Attention U-Net, likely due to their **simpler architecture** which reduces overfitting and handles binary class boundaries more effectively.

Comparison with State-of-the-Art

- Although our models performed well, they did not surpass **state-of-the-art** architectures such as **DeepLabV3+** [Chen et al., 2018] and **PSPNet** [Zhao et al., 2017].
- These models leverage advanced techniques such as **Atrous Spatial Pyramid Pooling (ASPP)** and **global context aggregation**, which provide superior results in high-resolution image segmentation tasks.

Potential Improvements

- **Enhancing Model Performance**: Fine-tuning hyperparameters and exploring deeper architectures could yield better results.
- **Integrating Transformers**: The use of **Vision Transformers (ViTs)** or **Swin Transformers** could improve long-range dependencies and feature representation.
- **Post-processing Techniques**: Implementing **CRFs (Conditional Random Fields)** or **Refinement Networks** could enhance segmentation smoothness.
- **Real-World Applications**: Extending this work to **medical imaging**, **autonomous driving**, and **satellite image analysis** would be beneficial.

Final Thoughts

- **Deep learning-based semantic segmentation** has vast potential in numerous fields.
- This study provides insights into the **strengths and limitations** of different U-Net architectures for segmentation tasks.

Thank you!

References

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018).

Encoder-decoder with atrous separable convolution for semantic image segmentation.
Proceedings of the European Conference on Computer Vision (ECCV), (pp. 801–818).

Oktay, O., Schlemper, J., Folgoc, L. L., & Lee, M. (2018).

Attention u-net: Learning where to look for the pancreas.
Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Ronneberger, O., Fischer, P., & Brox, T. (2015).

U-net: Convolutional networks for biomedical image segmentation.
Medical Image Computing and Computer-Assisted Intervention (MICCAI), (pp. 234–241).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017).

Pyramid scene parsing network.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2881–2890).

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018).

Unet++: A nested u-net architecture for medical image segmentation.
Deep Learning in Medical Image Analysis.