Colonoscopy Polyp Segmentation Using Unet Variants

First Author
Chu Quang Tung
20213594
tung.cq213594@sis.hust.edu.vn

Second Author
Dinh Viet Cuong
20213566
cuong.dv213566@sis.hust.edu.vn

Abstract

polyp detection and segmentation play a critical role in the early diagnosis and prevention of colorectal cancer. In this study, we propose a deep learning-based approach utilizing the UNet architecture for accurate polyp segmentation from colonoscopy images. The UNet model is particularly well-suited for medical image segmentation due to its encoder-decoder structure and skip connections that retain spatial information at different scales. Our method focuses on enhancing segmentation performance by addressing challenges such as variations in polyp size, shape, and texture. The proposed model is evaluated on publicly available polyp segmentation datasets, achieving promising results in terms of Dice coeffiency metrics. The findings demonstrate the effectiveness of the UNet model in improving the accuracy and reliability of automated polyp detection systems, which can support clinicians in real-time colonoscopy procedures, ultimately reducing the risk of missed polyps and improving patient outcomes.

1. Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer-related deaths worldwide. The early detection and removal of polyps during colonoscopy procedures significantly reduce the risk of CRC development. Polyps can be classified into two categories: benign and malignant. Benign polyps pose no immediate threat, while malignant polyps have the potential to become cancerous. The timely and accurate segmentation of these polyps from colonoscopy images is, therefore, a crucial step in improving the prognosis for patients. However, manual polyp segmentation by medical experts is time-consuming, subjective, and prone to variability between different observers. This necessitates the development of automated solutions to assist clinicians in detecting and segmenting polyps with greater consistency and efficiency.

Recent advancements in deep learning have made significant contributions to the field of medical image analysis, particularly in tasks such as classification, detection, and segmentation. In particular, convolutional neural networks (CNNs) have demonstrated remarkable performance in image segmentation tasks across various medical imaging modalities. Among the CNN-based architectures, U-Net has gained widespread popularity due to its ability to achieve high accuracy in segmenting objects of varying sizes and shapes while maintaining computational efficiency. The success of U-Net has led to the development of several extended architectures, such as U-Net++, which further enhance the model's ability to capture complex features in medical images.

Despite the progress in deep learning-based segmentation methods, challenges remain in achieving robust polyp segmentation due to the variability in polyp appearance, size, texture, and lighting conditions in colonoscopy images. Traditional handcrafted feature extraction methods fall short in addressing these challenges, making deep learning a more suitable approach for this task. The use of U-Net-based models has shown promise in addressing these issues by leveraging encoder-decoder architectures to capture both low-level and high-level image features.

This paper focuses on the implementation of a U-Net++ model for the task of colonoscopy polyp segmentation. The U-Net++ architecture, with its nested and dense skip connections, offers enhanced feature extraction capabilities compared to the traditional U-Net. Our implementation leverages pre-trained encoders and data augmentation techniques to improve generalization performance on unseen data. We also adopt a robust preprocessing pipeline to ensure accurate differentiation between malignant (red) and benign (green) polyps in colonoscopy images.

The remainder of this paper is organized as follows. In Section 2, we review existing methods for medical image segmentation, particularly focusing on polyp segmentation. Section 3 provides a detailed description of the dataset used in this study, including preprocessing and augmentation techniques. Section 4 outlines the technical details of our U-Net++ implementation, followed by the experimental setup and evaluation metrics, our results and discussion in

Section 5.

2. Related Work

Medical image segmentation has significantly evolved with the introduction of deep learning, particularly convolutional neural networks (CNNs). Traditional image processing techniques, which relied on handcrafted features, were limited in handling the complex variability of medical images. Deep learning approaches, especially fully convolutional networks (FCNs), have demonstrated superior performance in medical image segmentation by automatically learning features from data.

One of the most influential architectures in biomedical image segmentation is the U-Net, introduced by Ronneberger et al. in 2015 . The Unet follows a U-shaped architecture with a symmetric structure consisting of two main paths: a contracting path (encoder) and an expanding path (decoder). The contracting path captures high-level contextual features through a series of convolutional and max-pooling layers, progressively reducing the spatial resolution. The expanding path uses transposed convolutions to restore the spatial resolution and combines it with features from corresponding layers in the encoder through skip connections. These skip connections help preserve fine-grained details and improve segmentation accuracy. UNet effectively balances both global context and local precision, making it suitable for pixel-level classification tasks. Figure 1 Show the architecture of Unet

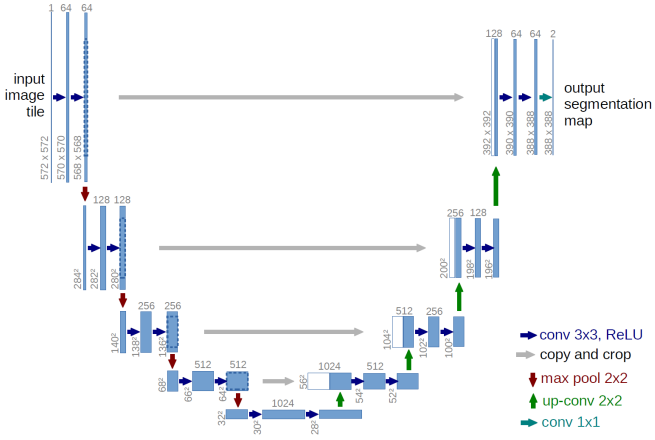

Figure 1. Unet Architecture

Following the success of U-Net, various extensions and improvements have been proposed. U-Net++ introduces nested skip pathways with intermediate convolutional layers between encoder and decoder. These nested pathways progressively refine feature maps, reducing the semantic gap between the two paths, and improving segmentation accuracy. While UNet is simpler and more computationally efficient, UNet++ offers better performance on complex segmentation tasks due to its more detailed feature fusion. However, the added complexity of UNet++ increases the computational cost of the model compared to the original UNet. Figure 2 Show the architecture of Unet++

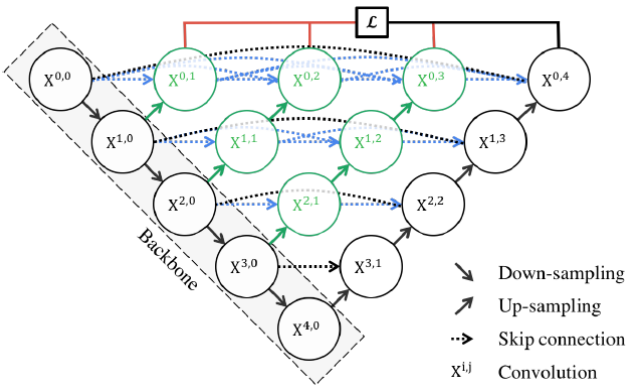

Figure 2. Unet Architecture

In the context of polyp segmentation from colonoscopy images, several studies have adopted U-Net-based architectures to achieve state-of-the-art results. The primary challenge in this domain is the variability in polyp size, shape, and texture, as well as the presence of artifacts and noise in the images. Data augmentation, particularly elastic deformations, has been crucial in improving the generalization ability of models trained on limited datasets.

This paper builds upon the U-Net architecture by adopting the U-Net++ variant, leveraging its enhanced feature extraction capabilities to improve polyp segmentation accuracy. The use of pre-trained encoders and advanced data augmentation techniques further enhances the model's robustness, making it suitable for real-world clinical applications.

3. Dataset

In this study, we utilize the BKAI-IGH-NeoPolyp dataset to evaluate the performance of our proposed UNet-based model for colonoscopy polyp segmentation. This dataset contains colonoscopy images and their corresponding ground truth segmentation masks. The dataset is divided into training and testing sets, with the training set comprising annotated images of polyps and the test set containing unannotated images for prediction.

3.1. Dataset Overview

The dataset consists of high-resolution colonoscopy images accompanied by pixel-wise annotated masks that mark the presence and boundaries of polyps. The annotations classify the pixels into three distinct classes:

- **Class 0: Background** —Non-polyp regions.
- **Class 1: Red Polyps** —Polyps identified as malignant regions, indicating potential cancerous growths.
- **Class 2: Green Polyps** —Polyps identified as benign regions, typically non-cancerous.

The multi-class segmentation task requires the model to distinguish between these classes accurately, making it more challenging and clinically relevant.

3.2. Dataset Structure

The dataset is divided into three subsets:

- **Training Set**: This subset is used to train the model. It includes images and corresponding ground truth masks.
- **Validation Set**: This subset is used during training to monitor the model's performance on unseen data and to prevent overfitting.
- **Test Set**: This subset is reserved for final model evaluation. It contains images without ground truth masks, and the model's performance is assessed based on its predictions.

The images vary in size and resolution. Therefore, consistent preprocessing is required to standardize the input dimensions.

3.3. Preprocessing Steps

To ensure uniformity in the input data, all images were resized to a resolution of 256x256 pixels for Resnet50 Encoder and 224x244 for MobilenetV2 Encoder. The resizing was necessary to achieve consistent batch sizes during training and to optimize the model's computational efficiency. Additionally, the images were converted from BGR to RGB color format, which is standard for most deep learning frameworks.

The corresponding masks were resized to the same dimensions and processed to ensure that the pixel values were correctly mapped to the respective classes. Specifically, the masks were transformed into binary representations, with each class assigned a unique integer value. Red masks indicating malignant polyps were mapped to Class 1, while green masks indicating benign polyps were mapped to Class 2.

3.4. Data Augmentation

Data augmentation techniques were applied to the training set to enhance the model's generalization capability. These augmentations simulate real-world variations in lighting, camera angles, and other factors that may affect image quality during colonoscopy procedures. The following augmentation techniques were used:

- **Horizontal and Vertical Flips**: Randomly flipping the images horizontally or vertically to introduce variations in perspective.
- **Random Gamma Adjustments**: Altering the brightness and contrast of the images to account for different lighting conditions.
- **RGB Shifts**: Shifting the red, green, and blue channels of the images to simulate variations in color balance.

These augmentations increase the diversity of the training data, making the model more robust to variations in real-world scenarios.

3.5. Ground Truth Masks

The ground truth masks provided in the dataset are essential for supervised learning. Each mask is a binary image where pixel values indicate the class of each pixel. The masks are generated using a combination of manual annotation and validation by medical experts to ensure accuracy.

The masks are processed to distinguish between the three classes mentioned earlier. The red and green polyp regions are separated from the background using thresholding techniques in the HSV color space. This approach helps in accurately identifying polyp boundaries, which is critical for precise segmentation. The red regions correspond to potentially malignant polyps, requiring careful attention during clinical diagnosis, while the green regions indicate benign polyps.

3.6. Test Set Evaluation

The test set contains images without corresponding ground truth masks. The model's predictions on this set are submitted for external evaluation using Dice coefficient.

Overall, the dataset provides a comprehensive collection of colonoscopy images with detailed annotations, enabling the development of a deep learning-based polyp segmentation model capable of identifying polyps with high accuracy and clinical relevance.

4. Implementation Details

The implementation of the polyp segmentation model using the UNet architecture was carried out using the PyTorch framework and the segmentation model library (smp). The pipeline includes dataset preparation, model architecture selection, training, evaluation, and prediction stages. The following sections provide detailed insights into each step of the implementation process.

4.1. Dataset Loading and Preprocessing

The data set utilized in this study consists of colonoscopy images and their corresponding mask annotations indicating the presence of polyps. The images were loaded from the specified directories and resized to 256x256 pixels to standardize the input size. The Dataset class was implemented to handle image reading, mask generation, and preprocessing.

The mask generation process involved converting the original mask images into binary masks using OpenCV. The masks were processed in the HSV color space to distinguish between red and green regions, corresponding to different polyp types. The red regions were identified using two sets of HSV thresholds to account for variations in hue. The green regions were identified similarly using another set of thresholds. These regions were then combined into a single mask with three classes: background, red polyp, and green polyp.

The data set was further divided into training and validation sets in an 80:20 ratio. Data augmentation techniques were applied to the training set to enhance the robustness of the model. These augmentations included horizontal and vertical flips, random gamma adjustments, and RGB shifts. The Albumentations library was used for this purpose, ensuring efficient and consistent transformations.

4.2. Model Architecture and Training

The UNet++ model was selected for the segmentation task, utilizing ResNet-50 and MobilenetV2 encoder pre-trained on the ImageNet dataset. The model was configured to accept three-channel RGB input images and three-class output segmentation masks.

The training process involved defining a custom data loader using the PyTorch DataLoader class. The batch size was set to 8, and both the training and validation data loaders were created using the split datasets. The Adam optimizer was employed with a learning rate of 0.0001, and the loss function used was CrossEntropyLoss to handle the multi-class segmentation task.

The training loop was structured to perform forward passes, calculate loss, backpropagate errors, and update the model parameters. Validation was conducted after each epoch to monitor the model's performance on unseen data. The best model weights were saved based on the validation loss, ensuring that the most performant model was retained.

4.3. Evaluation and Prediction

The evaluation phase involved loading the saved model checkpoint and setting the model to evaluation mode. The test images were resized and preprocessed similarly to the training data before being passed through the model to generate segmentation masks.

The output masks were resized back to the original image dimensions and converted to RGB format for visualization. A custom color dictionary was used to map the class labels to specific colors, facilitating clear and intuitive mask overlays on the original images.

For submission to external evaluation platforms, the predicted masks were encoded using Run Length Encoding (RLE). The RLE format efficiently compresses binary masks into strings, reducing storage and transmission requirements. A custom function was implemented to perform this encoding, ensuring compatibility with the expected submission format.

4.4. Conclusion

The implementation pipeline detailed in this section highlights the steps taken to develop a robust polyp segmentation model using UNet++. The process included dataset preparation, augmentation, model training, and evaluation, all of which were tracked and optimized using state-of-the-art tools and techniques. The resulting model demonstrates the feasibility of using deep learning techniques for accurate polyp detection in colonoscopy images.

5. Result

In this section, we present the performance of our U-Net++ model in segmenting colonoscopy polyps. The evaluation includes both quantitative metrics and qualitative results to assess the accuracy and robustness of the model in segmenting benign and malignant polyps.

5.1. Evaluation Metric

The primary evaluation metric used in this study is the Dice coefficient, which measures the pixel-wise agreement between the predicted segmentation and the corresponding ground truth. The Dice coefficient is calculated using the following formula:

$$Dice = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{1}$$

where X represents the predicted set of pixels, and Y represents the ground truth set of pixels. The Dice coefficient ranges from 0 to 1, where a value of 1 indicates perfect overlap between the prediction and the ground truth. In cases where both X and Y are empty, the Dice coefficient is defined to be 1.

The leaderboard score is computed as the mean Dice coefficient across all test samples. In this task, we classify polyps into two categories:

- **Class 0**: Healthy region.
- **Class 1**: Neoplastic polyps (malignant).
- **Class 2**: Non-neoplastic polyps (benign).

5.2. Results

The performance of the model was evaluated using the Dice coefficient, this metrics provide a comprehensive view of the model's ability to accurately identify polyp regions and distinguish between benign and malignant polyps. Table below compares the Dice score of our proposed model and other models.

Model	Dice Coefficient
Unet++ with Resnet50 Encoder	0.74620
Unet with Resnet34 Encoder	0.73228
Unet with mobilenetV2 Encoder	0.67029

As seen here, our proposed model slightly outperforms other models in terms of the Dice score.

Figure 3. Comparison of Train loss and evaluation loss of 200 epochs for each model

5.2.1. UNet++ vs UNet:

- UNet++ (both ResNet and MobileNet encoders) outperforms the default UNet in both training and validation loss.
- This is due to the nested skip connections in UNet++, which refine the feature maps and help reduce the semantic gap between the encoder and decoder.

5.2.2. Encoders Comparison:

- MobileNet encoder results in the most stable and low loss values, likely due to its lightweight architecture and strong generalization capabilities.
- ResNet encoder provides rich features but requires careful regularization to avoid overfitting.
- Default encoder performs poorly, as it may not extract meaningful features for complex segmentation tasks.

To further validate the performance of the model, we visualized the predicted segmentation masks overlayed on the original colonoscopy images.Figure 4 shows sample predictions from the test set The visual results indicate that the U-Net++ model can accurately segment polyps of varying sizes and shapes. The model effectively distinguishes between benign and malignant polyps, as evidenced by the correct identification of green and red regions in the masks.

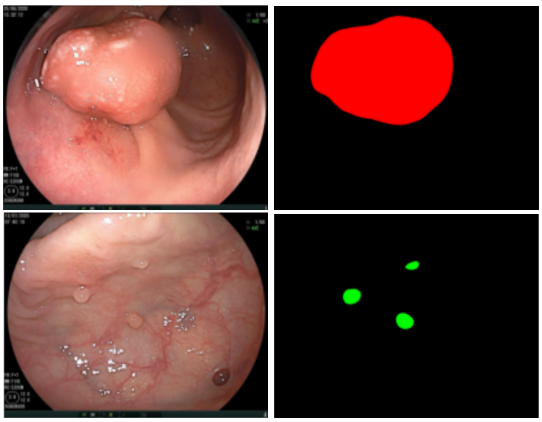

Figure 4. Sample predictions from the test set. The model's predictions align closely with the ground truth masks.

6. Future Work

Future research can explore several avenues to enhance the accuracy and robustness of colonoscopy polyp segmentation using UNet-based architectures. First, integrating attention mechanisms into UNet++ could further improve the model's ability to focus on relevant regions of interest, reducing false positives and false negatives. Second, the use of multi-scale feature extraction techniques can be investigated to handle polyps of varying sizes more effectively. Additionally, employing transformer-based encoders could improve the model's capability to capture long-range dependencies within colonoscopy images. Another promising direction is the use of semi-supervised or unsupervised learning methods to reduce reliance on large annotated datasets, which are often time-consuming to create. Finally, deploying these models in real-time clinical settings and validating their performance in diverse datasets from different institutions would ensure their generalizability and practical applicability in real-world scenarios.

7. Reference

[1] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018. https://arxiv.org/abs/1807.10165

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox.

”U-net: Convolutional networks for biomedical image segmentation.” Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015. https://doi.org/10.48550/arXiv.1505.04597 [3] Huy, Nguyen Sinh, et al. ”Polyp segmentation on colonoscopy image using improved Unet and transfer learning.” Journal of Military Science and Technology CSCE6 (2022): 41-55. https://doi.org/10.54939/1859-1043.j.mst.CSCE6.2022.41-55