CS482/682 Final Project Report Group 14
Benchmarking State-of-the-art U-Net Variants Performance on Surgical Tool Segmentation*

Chujian ”Harry” Yu[1], Zichen ”Frank” Fu[1], Xu ”Lance” Lian[1], and Le ”Chris” Wang[1]

[1]Department of Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, USA
{cyu62, zfu16, xlian3, lwang178}@jhu.edu

1 Introduction

Numerous architectures have been proposed in the literature to address problems of surgical tool segmentation. U-Net [8], an encoder-decoder-based architecture, has been hugely successful because of its integration of both local and global features. However, U-Net has disadvantages such as semantic gaps [7], causing more noise after concatenating shallow features with deeper features. Therefore, Multiple variants of U-Net have since then been proposed to improve on U-Net. This project aims to benchmark these different U-Net variants that could theoretically bridge the semantic gaps and attempt to further improve the segmentation accuracy by combining mechanisms of those variants.

We use the dataset released from the EndoVis2018 challenge [1] to train and test all our models. It contains 12 classes of objects in the surgical scene, as summarized in Table A in the Appendix. As a baseline, we have implemented a simplified U-Net with 3 max-pooling layers with intermediate batch-normalization layers and 3 upsampling layers. We have investigated several state-of-art models for Image segmentation, including U-Net++ [10], Attention U-Net [6], R2U-Net [9], and U-Net3+ [4]. These models are described in the Appendix. We present our methodology in Section 2. The results of the benchmarking are summarized in Section 3, and we discuss these results in Section 4.

2 Methods

Each frame of images in our dataset consists of a stereo pair of resolution 1280 × 1024. Since only the left image in each pair is annotated, we only use left frames for supervised learning. We plan to use the right frame images for pre-training with recolorization, but due to the limited time and computational resources we have, we decide to train our models without pre-training. We downsample the images to 320 × 256 to save memories. The images from the original dataset is divided into 20 sequences such that each sequence shares common surgery procedures and objects present. We merge 16 sequences and randomly divide the images into the training dataset and validation dataset with a ratio of 8:1. We pick the remaining 4 entirely unseen sequences as the test dataset. We set the training and validation batch size to 24 (reduced to 12 for Attention R2UNet and UNet3+ family to avoid GPU memory overflow). We use the Adam optimizer with a learning rate of 0.001, and DICE loss as the loss function. We use batch normalization and data augmentation as regularization to avoid overfitting. The augmentation we adopt includes random flipping (horizontal and vertical), random rotation, and affine transformation. We initialize each model with He initialization (all with random seed = 256) and train them over 20 epochs to ensure convergence. The model with the lowest validation loss is selected to be the best for test dataset evaluation.

Architectures: All the models are trained with the same depth for encoding and decoding (d = 4); the smallest scale of feature maps has a channel size of 512, and the parameter size is limited to around 10 million. We test for Nested UNet, UNet++, Attention UNet (AUNet), Attention R2UNet (AR2UNet), UNet3+, Attention UNet3+ (AUNet3+), UNet3+ with Atrous Depthwise Separable Convolution (ADSC UNet3+).

Metrics: To assess the performance of different models, we use the five popular metrics below to test our architectures: DICE score, IoU, Hausdorff Distance, Boundary displacement error, Recall. We describe these metrics in the Appendix.

[*] The codes developed in this project is publicly accessible at https://github.com/chujianyu/mldl2023project.git.

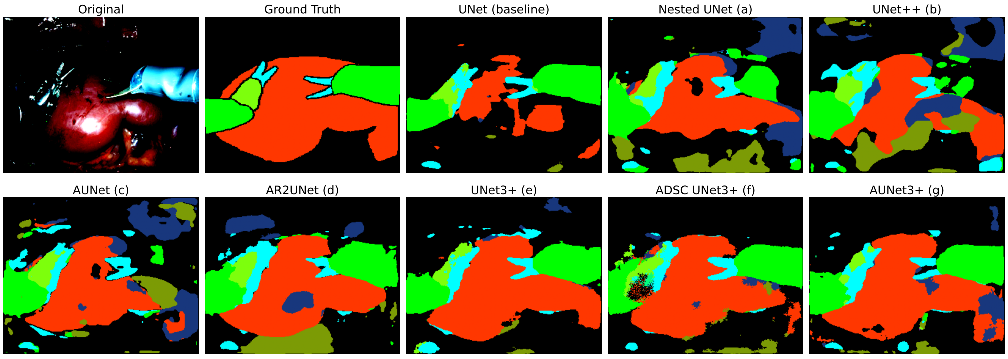

Figure 1: Performance of our trained models on unseen test data.

Model	**DICE**	**IoU**	**Hausdorff**	**Boundary**	**Recall**
UNet	0.6050	0.4233	**4.307**	19.43	0.5948
NestedUNet	0.5936	0.4125	4.769	17.54	0.5841
UNet++	0.6381	0.4638	4.372	16.93	0.6338
AUNet	0.6142	0.4329	4.314	17.32	0.6043
AR2UNet	0.6330	0.4253	4.455	17.26	0.5968
UNet3+	**0.6525**	0.5045	4.311	**15.13**	0.6707
AUNet3+	0.6262	**0.5209**	4.322	15.28	**0.6850**
ADSCUNet3+	0.6497	0.4446	4.363	16.21	0.6155

Table 1: Model Performance Metrics

3 Results

We illustrate the segmentation performance of the models we trained on one particular image in the test dataset in Fig. 1. Prediction from each model is shown to compare with the input image and the ground truth.

The results of each model based on the previously mentioned metrics are detailed in Table 1.

Nested UNet shows worse performance relative to other UNet variants; nevertheless, its performance significantly improves upon implementing Deep Supervision in UNet++.

UNet3+ achieves the highest dice score, second-lowest Hausdorff distance, and lowest boundary displacement error among all tested models.

ADSC UNet3+ achieves the second-highest dice score among all tested models. However, it showes an increased Hausdorff distance and boundary displacement error compared with UNet3+.

AUNet3+ achieves the highest recall and IoU among all of our tested models. It also achieves the second-lowest boundary displacement error, and a dice score above baseline but lower than UNet3+.

4 Discussion

U-Net++: UNet++ performs better compared to Nested UNet, indicating the advantage of Deep Supervision of the UNet architecture. The implementation of Deep Supervision compensates for potential data limitations in fine-tuning the intricacies of nested subnetworks. By calculating the loss with the uppermost nodes across all semantic levels, we enhance gradient flow, thereby optimally training each subnetwork.

AUNet: Attention mechanisms become more adaptive to different contexts within an image [6]. This adaptability aids in capturing intricate details and structures present in various parts of an image with a specifically trained weight, thereby improving the model's recall score.

AR2UNet: The residual-recurrent mechanism enhances this model's ability to memorize the features of previous feature maps. Therefore, compared to AUNet, AR2UNet segments less with isolated pixels and fragmented parts. However, this mechanism significantly increases the amount of parameters and training time.

U-Net3+: This model's full-scale skip connections effectively merge detailed spatial information from shallow layers with deep-layer semantics, significantly improving overall segmentation accuracy [4]. The lowest boundary displacement error of UNet3+ (which can be sensitive to poor predicted boundaries of small objects) suggests that full-scale skip connections are effective in capturing surgical tools at different scales [5].

ADSC UNet3+: The depthwise separable convolutions reduce the parameter size from 11 million to 1.4 million, making it suitable for fast inference and portable devices. Setting the dilation rate to 2 (vs.1) leads to an improved dice score, suggesting that atrous convolutions enhance the model's capability to capture global context, which is crucial for surgical tool segmentation [3]. The sparse, atrous sampling and depthwise separable convolutions, though efficient, may lead to a loss in spatial and cross-channel information. Future design may perform atrous convolution at multiple scales to capture both fine-grained details and global context.

AUNet3+: By using a more semantically-rich feature map from a deeper encoder layer as the gate signal to refine the more spatially-rich feature map, the attention blocks in AUNet3+ help the model focus on positive, pertinent regions, as evidenced by its highest recall. Because the surgical tools are usually confined in a specific area within the image, the attention mechanism effectively emphasizes these regions [6]. Compared with UNet3+, AUNet3+ shows a marginally higher Hausdorff distance and lower dice score, suggesting a decrease in total segmentation extent and preciseness in segmentation boundaries. Overall, despite a modest limitation in precise boundary definition, AUNet3+ exhibits enhanced selectivity and accuracy for positive samples.

Improvement Reasoning: We use the regular UNet as the baseline for benchmarking and then add extra mechanisms that we believe may improve the models' performance. By comparing the performance of each single additional mechanism (UNet vs. UNet++ vs. AUNet vs. UNet3+) and (AUNet vs. AR2UNet), we find that the full-scale skip connection featured by UNet3+ shows the most prominent improvement in terms of DICE score. By comparing the performance between (UNet vs. AUNet), we find the introduction of attention blocks shows consistent improvement in the model in terms of Recall. Therefore, we combine the mechanism of UNet3+ and attention, which we hypothesize could have an improved performance on our newly developed AUNet3+.

Observation: As shown in Fig. 1, the red label corresponds to ”kidney-parenchyma” (class 4) and the dark blue label corresponds to ”covered-kidney” (class 5). Multiple models give a false-positive dark-blue segmentation that contaminates the red part of the true label. However, since the categories represented by red and dark blue have close semantic meanings and spatial relevance, there is a high possibility that the models misrecognize them due to their similar representations in higher-dimensional space.

Future Development: This experiment is limited by the computational resources and thus not every possible combination of promising mechanisms was tested. For future development, the combination of R2Unet, Attention, atrous convolution, and UNet3+ is worthy of testing. Also, the depth and parameter size of the models can be further fine-tuned to optimally match the complexity of this segmentation task.

References

[1] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., Kori, A., Alex, V., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C.-S., Chiang, H.-A., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Kim, M., Kim, C., Kim, C., Kim, H., Lee, G., Ullah, I., Luna, M., Park, S. H., Azizian, M., Stoyanov, D., Maier-Hein, L., and Speidel, S. 2018 Robotic Scene Segmentation Challenge. *arXiv e-prints* (Jan. 2020), arXiv:2001.11190.

[2] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., Kori, A., Alex, V., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C.-S., Chiang, H.-A., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Kim, M., Kim, C., Kim, C., Kim, H., Lee, G., Ullah, I., Luna, M., Park, S. H., Azizian, M., Stoyanov, D., Maier-Hein, L., and Speidel, S. 2018 Robotic Scene Segmentation Challenge. *arXiv e-prints* (Jan. 2020), arXiv:2001.11190.

[3] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation. *ArXiv abs/1706.05587* (2017).

[4] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *arXiv e-prints* (Apr. 2020), arXiv:2004.08790.

[5] Mittal, H., Pandey, A. C., Saraswat, M., Kumar, S., Pal, R., and Modwel, G. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications 81*, 24 (2022), 35001–35026.

[6] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Y Hammerla, N., Kainz, B., Glocker, B., and Rueckert, D. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv e-prints* (Apr. 2018), arXiv:1804.03999.

[7] Pang, Y., Li, Y., Shen, J., and Shao, L. Towards bridging semantic gap to improve semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4230–4239.

[8] Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints* (May 2015), arXiv:1505.04597.

[9] Zahangir Alom, M., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *arXiv e-prints* (Feb. 2018), arXiv:1802.06955.

[10] Zhou, Z., Mahfuzur Rahman Siddiquee, M., Tajbakhsh, N., and Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *arXiv e-prints* (July 2018), arXiv:1807.10165.

Appendix

0	background-tissue
1	instrument-shaft
2	instrument-clasper
3	instrument-wrist
4	kidney-parenchyma
5	covered-kidney
6	thread
7	clamps
8	sulturing-needle
9	suction-instrument
10	small-intestine
11	ultrasound-probe

Table A: Objects to be Segmented in EndoVis2018

Architecture Description

The architectures benchmarked in this project are summarized as follows:

a. NestedUnet: Intermediate nodes employ a dense and nested approach to aggregate weighted inputs from all preceding nodes at the same level and a previous node from next level, feeding them into the decoding layers.

b. U-Net++: Improved Nested Unet using Deep Supervision to integrate all feature maps at the uppermost layer.

c. Attention U-Net (AUNet): The attention blocks are used here to assign a specific weight to different image parts at the skip connection. Activations from the irrelevant regions would be suppressed by this weight matrix after training in the skip connection.

d. Attention R2U-Net (AR2UNet): Residual recurrent blocks are used over the encoding and decoding layers with t-time-step recurrent convolutional layers and a residual path concatenates the raw input with the recurrented results.

e. U-Net3+: A full-scale skip connection is used, with inter-connnections between the decoders and intra-connections between the encoders and decoders, incorporating low-level details with high-level semantics from feature maps in different scales.

f. U-Net3+ with Atrous Depthwise Separable Convolution (ADSC UNet3+): We proposed a new model that replaces the convolutional layers of UNet3+ with Atrous Depthwise Separable Convolutions at a dilation rate of 2.

g. U-Net3+ with attention (AUNet3+): We proposed a new model that uses a full-scale skip connection with attention blocks that assign and train a specific weight for the original connections.

Adopted Metrics

We use the following metrics to evaluate our models:

a. DICE Score: Measures similarity by calculating the overlap between prediction and true segments.

$$DICE = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

b. IoU: The ratio of the overlap and union of the prediction and true segments.

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

c. Hausdorff Distance: Evaluates the maximum distance of mismatch between the boundaries of the predicted and actual segments.

$$H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}$$

d. Boundary displacement error: Calculates the average distance between t.he boundaries of the predicted and actual segments.

$$d(p_i, B_2) = \min_{p \in B_2} \|p_i - p\|_2$$

$$BDE = \frac{1}{2} \left(\frac{1}{|B_1|} \sum_{i=1}^{|B_1|} d(p_i, B_2) + \frac{1}{|B_2|} \sum_{j=1}^{|B_2|} d(p_j, B_1) \right),$$

where $B1$ and $B2$ are sets of edge pixels.

e. Recall: Proportion of correctly predicted positive observations to the total positives in true segmentation.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Main Proposal

A Problem Statement:

As surgical robots develop rapidly, the precision in their endoscopic working scenario stands out as one of the major challenges. To avoid touching the unexpected position during the surgery, it is important to correctly recognize and segment the robot's tools and the patient's organs. However, as the most commonly used model, U-Net has disadvantages such as semantic gaps[7], causing more noise after concatenating shallow features with deeper features. Therefore, the goal of this project is to benchmark U-Net variants from traditional models that could bridge the semantic gaps and add more modifications in attempt to further improve the accuracy.

B Related Work:

We have investigated several state-of-art models for Image segmentation, including U-Net++[10], Attention U-Net[6] and R2U-Net[9]. Each model has its advantages and disadvantages:

1. U-Net++: A deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways.

2. Attention U-Net: A novel attention gate (AG) model for medical imaging that automatically learns to focus on target structures of varying shapes and sizes.

3. R2U-Net: A recurrent residual convolutional neural network based on U-Net which is designed to capture fine-grained details by incorporating recurrent residual convolutional layers.

C Dataset:

For the purpose of benchmarking, we have selected the EndoVisSub 2018-Robotic Scene Segmentation dataset[2]. This dataset comprises approximately 2,400 instances of images captured during surgical procedures, each meticulously annotated for segmentation purposes. The preprocessing step would include mapping the color-coded masks to class labels.

D Proposed Approach:

Since we are comparing the similarity between the prediction and the labels, we will use dice score to evaluate the performance of the models:

$$\text{Dice} = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{1}$$

We will train the standard U-Net model and use its performance as the baseline for the benchmarking. We will construct the models for U-Net++, Attention U-Net, and R2U-Net and train them to perform surgical tool segmentation. We will fine-tune the hyperparameters for each of the four models during training. We will then compare the dice score for all of the models.

According to the results from the benchmarking and our understanding on the models, we plan to either improve the model with the highest dice score among the four or tentatively train a new model that we believe may achieve even better performance. The potential improvements we plan to make include loss function modifications, data augmentation, and deep supervision.