

Assessing the performance of deep learning-based models for prostate cancer segmentation using uncertainty scores

Pablo Cesar Quihui-Rubio¹, Daniel Flores-Araiza¹, Gilberto Ochoa-Ruiz¹,
Miguel Gonzalez-Mendoza¹, and Christian Mata^{2,3}

¹ Tecnológico de Monterrey, School of Engineering and Sciences, Mexico.

² Universitat Politècnica de Catalunya, 08019 Barcelona. Catalonia, Spain.

³ Pediatric Computational Imaging Research Group, Hospital Sant Joan de Déu, Esplugues de Llobregat, 08950, Catalonia, Spain

Abstract. This study focuses on comparing deep learning methods for the segmentation and quantification of uncertainty in prostate segmentation from MRI images. The aim is to improve the workflow of prostate cancer detection and diagnosis. Seven different U-Net-based architectures, augmented with Monte-Carlo dropout, are evaluated for automatic segmentation of the central zone, peripheral zone, transition zone, and tumor, with uncertainty estimation. The top-performing model in this study is the Attention R2U-Net, achieving a mean Intersection over Union (IoU) of $76.3\% \pm 0.003$ and Dice Similarity Coefficient (DSC) of $85\% \pm 0.003$ for segmenting all zones. Additionally, Attention R2U-Net exhibits the lowest uncertainty values, particularly in the boundaries of the transition zone and tumor, when compared to the other models.

Keywords: Segmentation · Uncertainty Quantification · Prostate · Cancer · Deep Learning · Computer Vision.

1 Introduction

Prostate cancer (PCa) is the most common solid non-cutaneous cancer in men and is among the most common causes of cancer-related deaths in 13 regions of the world [1]. According to a recent overview, in 2020 prostate cancer was the most frequently diagnosed cancer in males in 12 regions of the world, which translates to around 1.41 million new cases [1]. However, when detected in early stages, the survival rate for regional PCa is almost 100%. In contrast, the survival rate when the cancer is spread to other parts of the body is of only 30% [2].

Magnetic Resonance Imaging (MRI) is the most widely available non-invasive and sensitive tool for detection, localization and staging of PCa, due to its high resolution, excellent spontaneous contrast of soft tissues, and the possibility of multi-planar and multi-parametric scanning [3]. MRI can be also be used for PCa detection through the segmentation of Regions of Interest (ROI). The use of image segmentation for PCa can help determine the localization and the volume of the cancerous tissue [4]. Although prostate image segmentation is a relatively old problem and some novel methods have

been proposed, radiologists still perform a manual segmentation of the prostate gland and regions of interest (central zone, peripheral zone, and transition zone) [5]. This manual process is time-consuming, and is sensitive to the specialist experience, resulting in a significant intra- and inter-specialist variability. Therefore, automating the process of segmentation of prostate and gland regions of interest, may help save time for practitioner radiologists and additionally can be used as a training tool for others. One of the most popular architectures is the U-Net [6] model, which has been the inspiration behind many recent works in the literature, such as Swin U-Net [7], or R2U-Net [8]. While these models have yielded positive outcomes, inconsistencies in performance have been observed in U-Net-based segmentation due to the prostate’s anatomical structure. The boundaries between zones can distort semantic features, leading to unreliable results. Furthermore, automatic segmentation typically produces deterministic segmentation outcomes [9], and there is insufficient information available about the model’s confidence level [10]. Despite their successes in many medical image analysis applications, DL algorithms are usually not translated into real-world clinical scenarios because these do not provide information about the uncertainty associated with their prediction. This is problematic in the challenging context of pathological structures segmentation (e.g, tumors) as even the top-performing methods are prone to errors, and due to the lack of uncertainty information, it results impossible tell apart different sorts of erroneous predictions.

Therefore, the overall segmentation workflow can be improved by providing the uncertainties of the model that could allow end-users (e.g, clinicians) to review and refine cases with high uncertainty.

In this work, we carry out a thorough assessment of automatic prostate zone segmentation models using U-Net, Attention U-Net, Dense U-Net, Attention Dense U-Net, R2U-Net, Attention R2U-Net, and Swin U-Net architectures. Additional to the segmentation task, we include the pixel-wise estimation of the uncertainty, which can be done by obtaining a probability distribution of the weights of the model. The zones evaluated in this work are the central zone (CZ), the peripheral zone (PZ), transition zone (TZ), and, in the case of a disease, the tumor zone (TUM), unlike previous works which only evaluate CZ and PZ [10].

This paper has five sections including this introduction. Section 2 provides a review about what has been done in previous works related to prostate segmentation and uncertainty quantification. Section 3 the dataset used is described, followed by a description of the uncertainty quantification procedure in this segmentation task. In section 4 the results of the experiments are discussed in detail. Finally the conclusion of this work is presented in Section 6.

2 Related Work

2.1 Deep Learning Segmentation

For segmentation, one of the best known models in the literature is the U-Net architecture [6], which is the base for many other novel models. The work from Zhu et al. [11] proposes a U-Net based network to segment the whole prostate gland, obtaining encouraging results (DSC of 0.885). Moreover, this architecture has served as the

inspiration for some variants that enhance the performance of the original model. One example is the work from Clark et al. [12] that presents a model that combines concepts from the U-Net and the inception architectures. Another example is the work presented by Oktay et al. [13], which proposes the addition of attention gates inside the original U-Net model with the intention of focusing on specific target structures. The addition of attention has served as base for other architectures such as Attention Dense U-Net [14], Attention R2U-Net [8], among others. Also, the introduction of Transformers in U-Net architectures is a novel approach for segmentation task that had demonstrated a good performance in biomedical images, such as Swin U-Net [7]. Despite this, during the course of this study, no other research was found that segmented the four zones discussed in this paper. Therefore, the number of studies that consider a third zone (TZ) is still limited, this is more likely because the most common datasets used are PROMISE-12 and the one from the PROSTATEx challenge, with only CZ and PZ. In addition to that, providing a value that quantifies the uncertainty of the predictions can improve the overall workflow since it could easily allow refining uncertain cases by human experts.

2.2 Uncertainty Quantification

The work from Theckel et al. [15] introduces a U-Net architecture with spatial dropout to measure the uncertainty related to the segmentation of macular degeneration, utilizing different sizes of input data. The work from Suman et al. [16] applied the uncertainty quantification problem to retinal imaging using a ResNet-based model, modified with standard random dropout layers before every convolutional block. The work from Liu et al. [10] proposes an automatic segmentation of the prostate zones and introduces a pixel-wise uncertainty estimator using a ResNet50 backbone with attention and dropout layers.

3 Materials and Methods

3.1 Dataset

The dataset used in the present work was provided by *Universidad Polit cnica de Catalu a* (UPC) in Barcelona, and Centre Hospitalaire de Dijon in France. The dataset consists of three-dimensional T2-weighted fast spin-echo (TR/TE/ETL: 3600 ms/ 143 ms/109, slice thickness:1.25 mm) images acquired with sub-millimeter pixel resolution in an oblique axial plane. The number of patients in the dataset are 19, with a total of 205 images with their corresponding annotation masks (of prostate zones) used as ground truth which were validated by experts using a dedicated tool [17].

The full dataset of 205 images, contains four different combination of zones, being: (CZ+PZ), (CZ+PZ+TZ), (CZ+PZ+Tumor), and (CZ+PZ+TZ+Tumor) with 73, 68, 23, and 41 images, respectively. For the purpose of this work, the dataset was divided in 85% for training and 15% for testing.

3.2 Uncertainty Estimation in Prostate Segmentation

Epistemic and aleatory uncertainties are the two major types of uncertainty that can be quantified. Epistemic uncertainty captures the uncertainty related to the models parameters caused by the lack of data, and, aleatory uncertainty captures the noise inherent in the input data [10]. The sum of both uncertainties forms the predictive uncertainty.

In this work, the uncertainty of seven different U-net-based models was measured in the test set. To approximate the inference of a model, Monte Carlo (MC) dropout of a hidden layer was performed. MC Dropout is a technique used in neural networks to incorporate uncertainty. It treats a network with dropped-out neurons as Monte Carlo samples from all possible combinations, approximating a Gaussian process [10,18]. The minimization of cross-entropy loss is similar to minimizing the divergence of the predicted distribution [16]. Using MC Dropout, pixel-wise epistemic uncertainty can be computed as a variational Bayesian inference problem [16]. During predictions or testing, dropout is also necessary. The main focus of this study is to investigate the predictive uncertainty of prostate segmentation, which can be quantified using the entropy of the predictive distribution [10].

3.3 Proposed Work

This work uses the original U-Net model and six U-Net extensions from the literature: Attention U-Net [13], Dense U-Net [19], Attention Dense U-Net [14], R2U-Net [8], Attention R2U-Net, and Swin U-Net [7]. These architectures had demonstrated great performance segmenting biomedical images, even some of them with public prostate’s datasets including CZ and PZ. However, unlike in other works, we proposed to compare the performance segmenting the three main zones of the prostate (CZ, PZ, and TZ) and a tumor tissue if it is present, using the dataset described in Section 3.1.

Before the final training, an hyperparameter tuning process using a stratified 5-Fold validation with the training set was carried out using the base U-Net model in order to obtain the optimal combination of data augmentation, learning rate and an approximation of epochs for training. The results demonstrated that including data augmentation in the training did not increase significantly the performance of the models. Therefore we decided to use the original dataset without data augmentation due to computational resources and time processing. The previously mentioned models were trained for 145 epochs, using Adam optimizer with a learning rate of $1e - 4$ and Categorical Cross-Entropy (CCE) loss function. The performance was evaluated using Dice Score (DSC) and Intersection over Union (IoU) as the main metrics.

4 Results and Discussion

4.1 Quantitative Results

Table 1 shows a summary of evaluation results of the seven studied architectures, in terms of two metrics (DSC and IoU) and loss value. In order to obtain these results, the evaluation of each model was performed $T = 50$ times, and due to the incorporation of MC Dropouts the results were different each time. Therefore, the average of all evaluations and prostate zones is reported with their corresponding standard deviation.

Table 1: Comparison of model performance in segmentation metrics and loss value. The metrics are denoted by upward (\uparrow) or downward (\downarrow) arrows, indicating the desired direction of values. Bold values highlighted in green represent the best score achieved among all models.

| Model | IoU \uparrow | DSC \uparrow | Loss \downarrow |
|-----------------------|-------------------------------------|-------------------------------------|---------------------------------------|
| U-Net | 0.676 ± 0.021 | 0.770 ± 0.021 | 0.0139 ± 0.0007 |
| Attention U-Net | 0.688 ± 0.011 | 0.781 ± 0.010 | 0.0132 ± 0.0003 |
| Swin U-Net | 0.725 ± 0.014 | 0.816 ± 0.014 | 0.0134 ± 0.0002 |
| Dense U-Net | 0.754 ± 0.004 | 0.846 ± 0.004 | 0.0146 ± 0.0003 |
| Attention Dense U-Net | 0.760 ± 0.006 | 0.847 ± 0.005 | 0.0154 ± 0.0004 |
| R2U-Net | 0.764 ± 0.002 | 0.850 ± 0.002 | 0.0119 ± 0.0001 |
| Attention R2U-Net | 0.763 ± 0.003 | 0.850 ± 0.003 | 0.0113 ± 0.0001 |

Based on the metrics values, it can be seen that U-Net was the model with worst performance. The use of attention to focus on the ROI helped to slightly outperform the performance in segmentation tasks compared to the original U-Net by around 1 – 2% for IoU and DSC.

Moving to Swin U-Net, a novel architecture from the state-of-the-art that uses Swin Transformers [7,?] achieved to increase the IoU and DSC values by more than 7%, and lower loss value compared to U-Net.

In the case of Dense U-Net, the performance of the model exceeds the previous three architectures, with IoU and DSC scores 11% and 10% better than the base U-Net, respectively, with a loss value of 0.0146. As a plus, this model did not need more computational resources or time during its training compared to base U-Net. The next model consisted on the incorporation of attention modules to Dense U-Net, which again outperformed all the previous models in the segmentation metrics by 12% of IoU, and 10% of DSC compared to U-Net. However, it achieved the higher loss value among all of 0.0154.

The last two architectures R2U-Net and Attention R2U-Net achieved very similar results, but outperformed all the other models with values of 76.4% and 85% for IoU and DSC, respectively, and the lowest loss value of 0.0113 for the Attention R2U-Net.

As mentioned before, an uncertainty comparison between the architectures was carried out per each prostate zone, as well as for the full image with its corresponding standard deviation as it is shown in Figure 1. The results shown in this figure can help us to determine, in relation with previous table, which model achieved to segment with more certain the prostate and its zones.

In Figure 1 it is observed that overall, the model that had the lowest mean uncertainty segmenting all the images in the test set was R2U-Net with a mean value of 0.048 ± 0.014 after 50 predictions, validating the results obtained in the Table 1, being the most reliable and accurate model overall thanks to the use of recurrent and residual units to get more context information.

Furthermore, the Attention U-Net was the one with the highest uncertainty overall with a value of 0.086 ± 0.023 , having poor results in comparison to the other models.

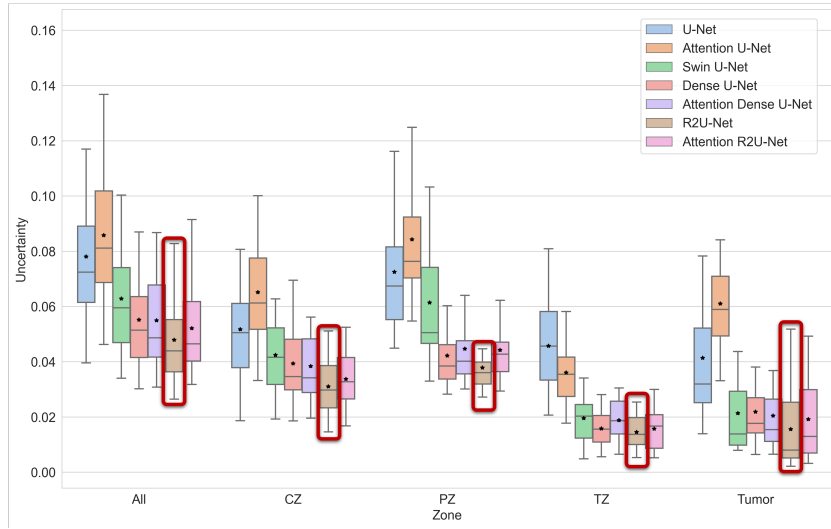


Fig. 1: Comparison of Uncertainty per each class between DL Architectures. The mean uncertainty could be identify with a black star inside each box, and the line represents the median uncertainty value obtained, the best model is indicated with a red box for each zone.

U-Net and Swin U-Net obtained very similar results in most of the prostate zones, although in the case of the TZ and Tumor, Swin U-Net achieved lower uncertainty.

Dense U-Net, Attention Dense U-Net and Attention R2U-Net succeeded in obtaining smaller uncertainty mean values than U-Net (0.055 ± 0.018 , 0.054 ± 0.018 , and 0.052 ± 0.014 , respectively). Although, TZ and Tumor are the zones less present in the dataset, and where it looks to be more complex to segment, models like R2U-Net and Attention R2U-Net managed to achieved a great segmentation performance and uncertainty values in average of those zones in the test set. It is important to notice that both results are correlated. These models managed to be adequately trained to perform the most accurate segmentation task among the others, which can give more confidence to radiologists when using a prostate segmentation tool based in this trained model.

4.2 Qualitative Results

In Figure 2, a qualitative comparison is presented among the predictions of each model using four different example images from the dataset. The comparison involves all possible combinations of zones. The first two columns display the original T2-MRI image of the prostate and its corresponding ground truth mask. Subsequently, each column represents the average of probabilities obtained from 50 predictions for each model. It can be observed that the first two zone combinations (Image A and B in Figure 2) are relatively easier for most models, as they produce segmentation that closely resemble the ground truth. However, certain models such as U-Net and Swin U-Net appear to misclassify pixels as TZ even when they are not present in the ground truth. Nevertheless, based on the examples in the test set, the models have been trained effectively to

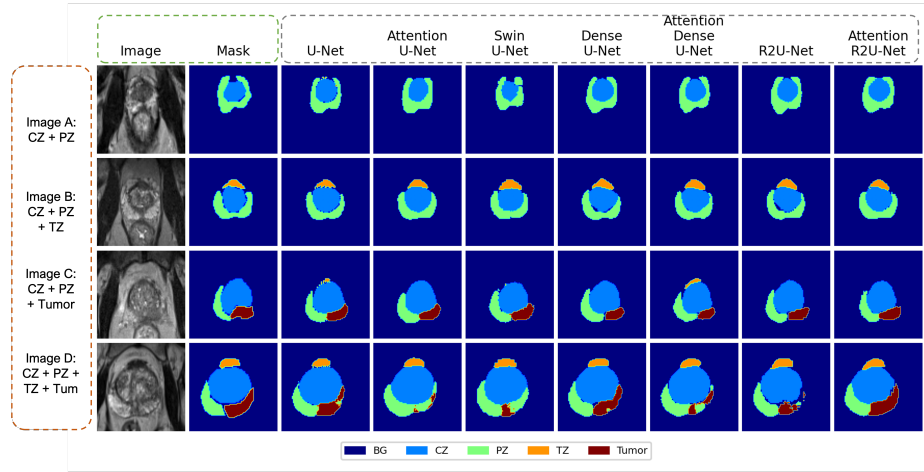


Fig. 2: Comparison of average segmentation after 50 predictions for each model in all the combinations of zones in the dataset.

achieve satisfactory segmentation performance on images containing CZ and PZ, and some including TZ.

Regarding the other two combinations that include the tumor, they posed the most complex segmentation challenge with notable variation among models. In Image C of Figure 2, models like U-Net and Attention Dense U-Net incorrectly classified a TZ region that was not identified in the ground truth. Meanwhile, other models tended to excessively smooth the original segmentation, yielding a seemingly good but possibly inaccurate result. However, when visually compared to the ground truth, the best segmentation in this example was achieved by R2U-Net and Attention R2U-Net.

For the last example, most models struggled to accurately segment the tumor. Surprisingly, U-Net and Dense U-Net produced acceptable results, but Attention R2U-Net demonstrated the best overall performance.

Figure 3 illustrates the significance of uncertainty by displaying the same four examples as in the previous figure, along with corresponding uncertainty maps represented as heat maps for each trained model. The temperature of the image indicates the level of uncertainty, with higher temperatures indicating greater uncertainty in those pixels, while lower temperatures indicate higher certainty in the model's pixel segmentation.

The model with the highest uncertainty, particularly around the boundaries of TZ and tumor, is U-Net, followed by Attention U-Net. This observation is evident. Furthermore, as previously mentioned, the first two examples were easier for the models, resulting in relatively low uncertainty across most of them. When dealing with images containing tumors, the inclusion of dense blocks enhanced model certainty. However, the utilization of recurrent residual blocks and attention modules surpassed other models, achieving acceptable predictions in the test set with low uncertainty values, even in TZ and tumor tissues.

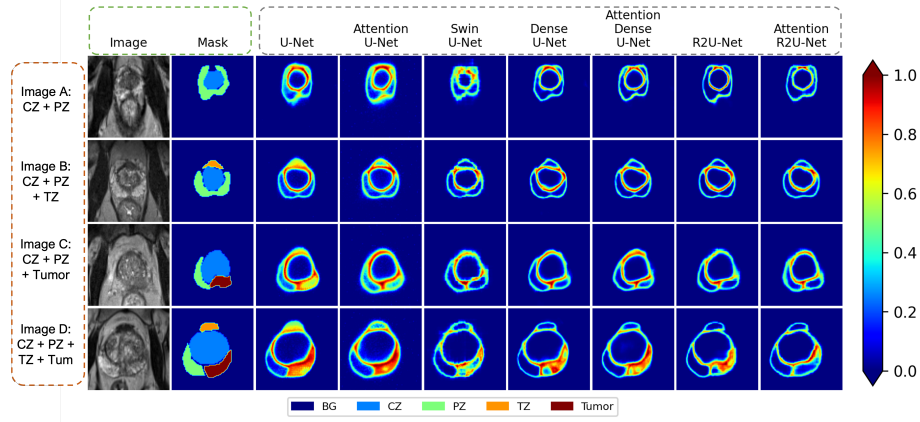


Fig. 3: Comparison of uncertainty maps after 50 predictions for each model with previous examples.

5 Application

In order to have a computer-aided tool which can be used for radiologists or clinicians, we proposed a Web App using Flask framework which we called '*ProstAI*', and it was designed to have easier access to predict images using the best trained model with MC dropouts: Attention R2U-Net. This app predicts the segmentation mask, as well as the uncertainty map, which is very helpful to indicate the experts which are the pixels where the model has higher uncertainty about their segmentation, an example is shown in Figure 4.

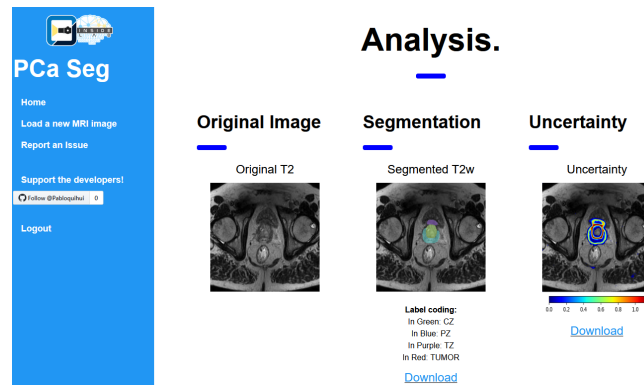


Fig. 4: Example of the analysis page of the '*ProstAI*' app using a prostate image from the Test set.

This tool is proposed for experimental usage, further information about the app and an example of usage can be found in: <https://github.com/pabloquihui/ProstAI>.

6 Conclusion

This study makes a valuable contribution to prostate cancer segmentation by introducing the segmentation of transition and tumor zones, along with the quantification of uncertainty, which has received limited attention in existing literature. The utilization of a private dataset validated by multiple experts, including two radiologists and two oncologists, enhances the reliability and accuracy of the findings. A comparison of seven different deep learning models was conducted using segmentation metrics, uncertainty scores, and visual inspection. Among these models, Attention R2U-Net emerged as the top-performing approach in both analyses. The inclusion of recurrent residual blocks in U-Net (R2U-Net) notably enhanced the segmentation results by capturing additional contextual information. Furthermore, Attention R2U-Net demonstrated exceptional proficiency in segmenting all prostate zones, exhibiting superior performance in metrics and yielding lower average uncertainty estimated using the MC method. This highlights the positive impact of attention modules on improving segmentation and, more significantly, reducing uncertainty in predictions by focusing on the ROI.

Moreover, a web app has been developed with a focus on experimental use for radiologists. This app provides more accurate, consistent, and faster results and displays the uncertainty map for each predicted image. The uncertainty map provides a visual representation of the pixels in which the model is uncertain about the segmentation, giving radiologists a better idea of the areas that require further analysis.

7 Acknowledgments

The authors wish to acknowledge the Mexican Council for Science and Technology (CONACYT) for the support in terms of postgraduate scholarships in this project, and the Data Science Hub at Tecnológico de Monterrey for their support on this project. This work has been supported by Azure Sponsorship credits granted by Microsoft's AI for Good Research Lab through the AI for Health program. The authors would also like to thank the financial support from Tecnológico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # E120 - EIC-GI06 - B-T3 - D.

References

1. Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. 1
2. AstraZeneca. A personalized approach in prostate cancer. <https://www.astrazeneca.com/our-therapy-areas/oncology/prostate-cancer.html>, May 2020. Accessed October 17, 2021. 1

3. M. Chen, H-D. Dang, J-Y. Wang, C. Zhou, S-Y. Li, W-C. Wang, W-F. Zhao, Z-H. Yang, C-Y. Zhong, and G-Z. Li. Prostate cancer detection: Comparison of t2-weighted imaging, diffusion-weighted imaging, proton magnetic resonance spectroscopic imaging, and the three techniques combined. *Acta Radiologica*, 49(5):602–610, 2008. [1](#)
4. R. Haralick and L. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985. [1](#)
5. N. Aldoj, F. Biavati, F. Michallek, S. Stober, and M. Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. *Scientific Reports*, 10, 08 2020. [2](#)
6. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. [2](#)
7. Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland. [2](#), [3](#), [4](#), [5](#)
8. Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *CoRR*, abs/1802.06955, 2018. [2](#), [3](#), [4](#)
9. Sadia Basar, Mushtaq Ali, Gilberto Ochoa-Ruiz, Mahdi Zareei, Abdul Waheed, and Awais Adnan. Unsupervised color image segmentation: A case of rgb histogram based k-means clustering initialization. *PLOS ONE*, 15(10):1–21, 10 2020. [2](#)
10. Yongkai Liu, Guang Yang, Melina Hosseiny, Afshin Azadikhah, Sohrab Afshari Mirak, Qi Miao, Steven S. Raman, and Kyunghyun Sung. Exploring uncertainty measures in bayesian deep attentive neural networks for prostate zonal segmentation. *IEEE Access*, 8:151817–151828, 2020. [2](#), [3](#), [4](#)
11. Qikui Zhu, Bo Du, Baris Turkbey, Peter L. Choyke, and Pingkun Yan. Deeply-supervised CNN for prostate segmentation. *CoRR*, abs/1703.07523, 2017. [2](#)
12. T. Clark, A. Wong, M. Haider, and F. Khalvati. Fully deep convolutional neural networks for segmentation of the prostate gland in diffusion-weighted mr images. pages 97–104, 06 2017. [3](#)
13. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. [3](#), [4](#)
14. Shuyi Li, Min Dong, Guangming Du, and Xiaomin Mu. Attention dense-u-net for automatic breast mass segmentation in digital mammogram. *IEEE Access*, 7:59037–59047, 2019. [3](#), [4](#)
15. Tinu Theckel Joy, Suman Sedai, and Rahil Garnavi. Analyzing epistemic and aleatoric uncertainty for drusen segmentation in optical coherence tomography images, 2021. [3](#)
16. Suman Sedai, Bhavna Antony, Dwarikanath Mahapatra, and Rahil Garnavi. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning, 2018. [3](#), [4](#)
17. C. Mata, J. Munuera, A. Lalande, G. Ochoa-Ruiz, and R. Benitez. Medicalseg: A medical gui application for image segmentation management. *Algorithms*, 15(06), 2022. [3](#)
18. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. [4](#)
19. Yufeng Wu, Jiachen Wu, Shangzhong Jin, Liangcai Cao, and Guofan Jin. Dense-u-net: Dense encoder–decoder network for holographic imaging of 3d particle fields. *Optics Communications*, 493:126970, 2021. [4](#)