Blurry Boundary Segmentation with Semantic-Aware Feature Learning

Qiuyu Xiao[1], Dong Nie[1,]

[1]Department of Computer Science, University of North Carolina at Chapel Hill, USA

Abstract. Encoder-decoder architectures are widely adopted for medical image segmentation tasks. These models utilize lateral skip connections to capture and fuse both semantic and resolution information in deep layers, enhancing segmentation accuracy. However, in many applications, such as images with blurry boundaries, these models often struggle to precisely locate complex boundaries and segment tiny isolated parts due to the fuzzy information passed through the skip connections from the encoder layers. To solve this challenging problem, we first analyze why simple skip connections are insufficient for accurately locating indistinct boundaries. Based on this analysis, we propose a semantic-guided encoder feature learning strategy. This strategy aims to learn high-resolution semantic encoder features, enabling more accurate localization of blurry boundaries and enhancing the network's ability to selectively learn discriminative features. Additionally, we further propose a soft contour constraint mechanism to model the blurry boundary detection. Experimental results on real clinical datasets demonstrate that our proposed method achieves state-of-the-art segmentation accuracy, particularly in regions with blurry boundaries. Further analysis confirms that our proposed network components significantly contribute to performance improvements. Experiments on additional datasets validate the generalization ability of our proposed method.

1 Introduction

Automatic image segmentation is an essential step in many medical image analysis applications, including computer-aided radiation therapy, disease diagnosis, and treatment effect evaluation. One of the major challenges in this task is the inherently blurry nature of medical images (e.g., CT, MR, and microscopic images), which often results in low-contrast and even vanishing boundaries, as illustrated in Fig. 1.

Many encoder-decoder based networks have been proposed for semantic segmentation [6, 9, 13, 14], achieving promising performance across various tasks. UNet [9], a typical encoder-decoder architecture that combines shallow and deep features through skip connections, is widely used in numerous image segmentation tasks. Several works have been proposed to enhance UNet [7, 10]. However, Heller et al. [2] found that while deep segmentation models are robust in non-boundary regions, they are less effective at handling boundaries. Specifically, these models often struggle to segment blurry boundaries, particularly in cases of extremely low tissue contrast. For example, prostate boundaries in MR or CT pelvic images are often blurry. To address this challenge, we argue that high-resolution, rich semantic feature learning is essential.

In addition to the variants of UNet, other approaches have been proposed to better delineate boundaries [4, 8, 11, 12, 15, 16]. Ravishankar et al. [8] introduced a multi-task network that robustly segments organs by jointly regressing boundaries and foreground. Zhu et al. [16] developed a boundary-weighted domain adaptive neural network to accurately extract MRI prostate boundaries. Lee et al. [4] a novel

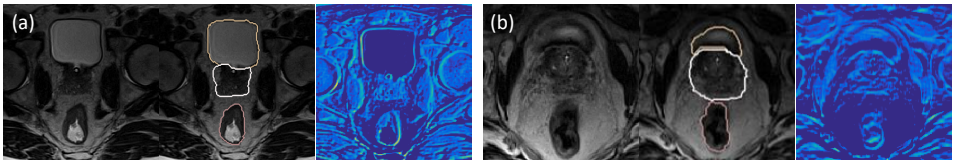

Fig. 1: Illustration of the blurry and vanishing boundaries within pelvic MRI images, together with overlaid ground truth contour and the typical feature maps in the encoder layer of a conventional UNet. (a) and (b) are the two typical slices of two subjects, in which boundaries of bladder and rectum are relatively clear, but prostate is blurry.

structure boundary preserving segmentation framework to tackle the ambiguous boundary delineation. However, these methods do not consider the high similarity of voxels around blurry boundaries. Consequently, directly classifying or regressing voxels as boundary or non-boundary is not ideal.

In this paper, we propose a novel semantic-guided encoder feature learning mechanism to enhance skip connections in existing encoder-decoder architectures, aiming to improve segmentation of low-contrast medical images. Our network design is based on the idea of explicitly utilizing high-resolution semantic information to address the inaccuracies in boundary delineation found in current encoder-decoder networks. Specifically, we concatenate low-layer (encoder) feature maps with high-layer (decoder) feature maps and employ both channel-wise and spatial-wise attention mechanisms. This approach facilitates the learning and selection of high-resolution semantic encoder feature maps.

With these improved encoder feature maps, we further concatenate (or element-wise add) them to the corresponding decoder layers within the encoder-decoder framework. Additionally, we propose using soft labels to indicate the probability of a voxel being on the boundary. Correspondingly, a soft cross-entropy loss is introduced as a metric to address the blurry boundary delineation problem.

2 Method

The architecture of our proposed framework is presented in Fig. 2. This framework utilizes an encoder-decoder architecture and addresses three tasks: segmentation, clear boundary detection, and blurry boundary detection. The proposed semantic-guided encoder feature learning module (SGM) is further highlighted in Fig. 3.

In the following subsections, we will analyze the deficiencies of skip connections in the current encoder-decoder framework. We then introduce the proposed semantic-guided encoder feature learning strategy. Additionally, we will describe the soft contour constraint for blurry boundary delineation. Finally, we will provide the implementation details.

2.1 Analysis of Skip Connection in Encoder-Decoder Architecture

In the classical encoder-decoder architecture [9], shallow and deep features are typically complementary. Shallow features are rich in resolution but lack sufficient semantic information, while deep features are semantically rich but lack spatial details. The skip connection proposed in UNet [9] aims to provide high-resolution information from the shallow (encoder) layers to the deep (decoder) layers, improving localization precision without sacrificing classification accuracy. However, the raw (simple) skip connection has several drawbacks: a) It introduces 'noise' (unnecessary information) to the deep layers, which negatively affects the concatenation of feature maps, as shown in the visualized encoder feature maps in Fig. 1. b) The

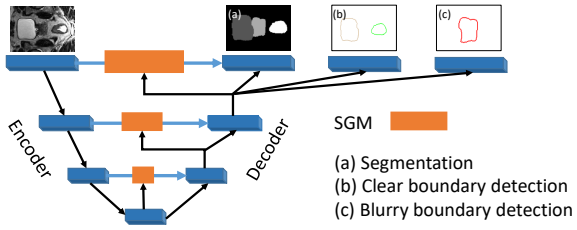

Fig. 2: Illustration of the architecture of our proposed method, which consists of a semantic-guided module (SGM). (a) means a segmentation branch, and (b) and (c) indicate boundary detection branches.

significant gap between shallow and deep features reduces the effectiveness of this combination. c) For clear boundaries (e.g., bladder and rectum), the encoder feature maps provide sufficiently precise localization information, allowing the raw skip connection to work effectively (as shown in Fig. 1). However, for blurry boundaries (e.g., prostate), the encoder feature maps fail to accurately describe these regions, resulting in poor localization with simple skip connections (as shown in Fig. 1). Therefore, it is crucial to select discriminative features rather than merely suppressing non-discriminative features from shallow layers. In other words, we need to learn high-resolution semantic features from the encoder. To achieve this, Roy et al. [10] proposed a concurrent spatial-and-channel squeeze and excitation module to enhance meaningful features and suppress weak ones. Oktay et al. [7] introduced a gated attention mechanism to select the salient parts of the feature maps, further improving UNet. However, in both works, the feature learning process is conducted in an *implicit* manner, which limits learning efficiency.

2.2 Semantic-guided Encoder Feature Learning

To overcome the aforementioned problems, we propose to *explicitly* learn high-resolution semantic features (which are also more discriminative) from shallow (encoder) layers with semantic guidance from deep (decoder) layers. The key idea is to encode semantic concepts from deep layer features to guide the learning of shallow features.

As shown in Fig. 3, our semantic-guided feature learning module (i.e., SG module or SGM) is designed to selectively enhance or suppress the features of shallow layer at each stage, thereby improving the consistency between shallow and deep layers without losing resolution information. In addition to the widely-used channel-wise encoder, we have also designed a spatial-wise encoder as described below.

We consider the feature maps of a certain encoder layer (i.e., shallow features) to be $S = \{s_1, s_2, ..., s_K\}$, where $s_i \in R^{H \times W \times T}$. We also assume the upsampled feature maps in the corresponding decoder layer (deep features) to be $D = \{d_1, d_2, ..., d_K\}$, where $d_i \in R^{H \times W \times T}$. We concatenate the two group of feature maps together and thus result in a bank of *high-resolution and rich-semantic mixed* feature maps as shown in Eq. 1.

$$F = \{s_1, s_2, ..., s_K, d_1, d_2, ..., d_K\}. \tag{1}$$

Channel-wise Encoding: With a global average pooling layer, we obtain a vector $Q = \{q_1, q_2, ..., q_K, ...q_{2K}\}$, where q_k is a scalar and corresponds to the averaging value of the k-th feature maps in F. Then, two successive fully connected layer are adopted to fuse the resolution and semantic information: $Z = W_1(\text{ReLU}(W_2 Q))$, with $W_1 \in R^{K \times K}$ and $W_2 \in R^{2K \times K}$. This encodes the channel-wise dependencies by considering both shallow and deep features. We apply a sigmoid activation

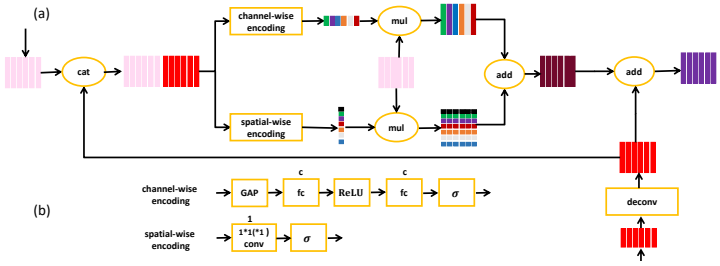

Fig. 3: Illustration of the proposed semantic-guided module (SGM), as shown in (a). The pink blocks represent the features of shallow layers, while the red ones represent the features of deep layers. Different from direct skip connection in UNet, we propose using semantic concept from deep layers to guide feature learning in the corresponding shallow layers, for which a channel-wise encoder and a spatial-wise encoder are both proposed, as shown in (b). 'GAP' means Global Average Pooling.

function to map the neurons to probabilities so that we can formulate as a channel-wise importance descriptor, which can be described as $\sigma(Z)$. Thus, the semantic-guided channel-wise encoded feature maps are formulated as Eq. 2.

$$SGCF = \{\sigma(z_1)\, s_1, \sigma(z_2)\, s_2, ..., \sigma(z_K)\, s_K\} \tag{2}$$

Note that the weight $\sigma(z_k)$ before the shallow feature map s_k can be viewed as an indicator of how important this specific feature map is. Thus, we argue this channel-wise encoding is actually a semantic-guided feature selection process in a channel-wise manner, which is able to ignore less meaningful feature maps and emphasize the meaningful ones. In other words, it can help remove the 'noise' and retain the useful information. More importantly, since $\sigma(Z)$ has taken both high resolution and rich semantic information into account, it has more discriminative capacity than the case of only considering shallow layer information in [10].

Spatial-wise Encoding: Now we come to consider the spatial-wise importance to achieve better fine-grained image segmentation.

Based on the concatenated feature maps F, we apply a $2K \times 1 \times 1 \times 1$ convolution to squeeze the channels. Therefore, we can obtain a one-channel output feature map U, where $U \in R^{H \times W \times T}$. We directly apply sigmoid function to acquire a probability map for U. Similarly, the semantic-guided spatial-wise encoded shallow feature maps can be described in Eq. 3.

$$SGSF = \{\sigma(U) \otimes s_1, \sigma(U) \otimes s_2, ..., \sigma(U) \otimes s_K\} \tag{3}$$

Since $\sigma(U_{h,w,t})$ corresponds to the relative importance of a spatial information at (h, w, t) of a given shallow layer feature map, it can help select more important features to relevant spatial locations and also ignore the irrelevant ones. Moreover, $\sigma(U)$ is a fusion of both resolution and rich semantic information, thus it can provide a better localization capacity even for the blurry boundary regions which cannot done by [10]. As a result, we view this spatial-wise encoding as a semantic-guided recalibration process.

Combination of Encoded Feature Maps: Now we can formulate both channel-wise and spatial-wise encoding by a simple element-wise addition operation, as shown in eq. 4.

$$SGF = \mathrm{SGCF} + \mathrm{SGSF} \tag{4}$$

This SGF considers both channel-wise encoded and spatial-wise encoded information, thus, it contains *not only* the discriminative (semantic) features, *but also* more accurate localization information.

Final Combination with Deep-Layer Feature Maps: To this end, we can simply complete the concatenation operation or element-wise addition operation. Instead of using the shallow feature maps S, we use the channel-wise and spatial-wise encoded shallow feature maps SGF to combine with the deep-layer feature maps D (through concatenation or element-wise addition). Compared with the raw skip connection in UNet, our encoded shallow feature maps SGF has same resolution but much more semantic and precise localization information (especially for the blurry regions), and thus can make the combination more reasonable. At the same time, since the operations in the encoder are mostly $1 \times 1 \times 1$ convolution, the number of parameters just increases a little bit.

To further increase the model's discriminative capacity, we also adopt the multi-scale deep supervision strategy as in [14] after feature fusion at each stage.

2.3 Boundary Delineation with Soft Contour Constraint

In mammal visual system [3], contour delineation closely correlates with object segmentation. To incorporate the knowledge to improve the segmentation accuracy, we integrate the task of contour detection with the task of segmentation, assuming that introducing a task of contour detection can help guide the network to concentrate more on the boundaries of organ regions, thus helping overcome the adverse effect of low tissue contrast. In this paper, as shown in Fig. 2, two boundary detection tasks are added to the end of the network as auxiliary guidance.

To extract the contour for training, we first delineate the boundaries of different organs by performing Canny detector on the ground-truth segmentation. For the organs with clear boundaries (i.e., bladder and rectum in our case), we model the problem as a classification problem. However, due to the potential sample imbalance problem, we propose using focal loss to alleviate such an issue, as shown in Eq. 5.

$$L_{cboundary} = -\sum_h \sum_w \sum_t \sum_{c \in csets} I_{\{Y_{h,w,t,c}\}} (1 - \hat{p}\,(X_{h,w,t}; \theta))^{\gamma}\, (1 - \hat{p}\,(X_{h,w,t}; \theta)) \tag{5}$$

Note that, for the regions with blurry boundaries (i.e., prostate in our case), the voxels near the boundaries look almost the same. As a result, it will be more reasonable to assign soft labels (instead of hard labels) around the ground-truth boundaries. Thus, we can formulate the blurry-boundary delineation task as a soft classification problem, which estimates the probability of each voxel being on the organ boundaries. Then, for these blurry boundaries, we further exert a Gaussian filter (with a bandwidth of δ, i.e., empirically set to 3 in our study) on the obtained boundary map. In other words, for each voxel, we generate an approximate probability belonging the blur boundary of an organ. Hence, we can formulate soft classification as a soft cross-entropy loss function as defined in Eq. 6.

$$L_{bboundary} = -\sum_h \sum_w \sum_t p_{h,w,t}\, (1 - \hat{p}\,(X_{h,w,t}; \theta)) \tag{6}$$

2.4 Implementation Details

Pytorch[1] is adopted to implement our proposed method shown in Fig. 2. The code can be obtained by this link[2]. We adopt Adam algorithm to optimize the network. The input size of the segmentation network is $144 \times 144 \times 16$. The network weights are initialized by the Xavier algorithm, and weight decay is set to be 1e-4. For the network biases, we initialize them to 0. The learning rate for the network is initialized to 2e-3, followed by decreasing the learning rate 10 times every 2 epochs during the training until 1e-7. Four Titan X GPUs are utilized to train the networks.

[1] https://github.com/pytorch/pytorch
[2] https://github.com/ginobilinie/SemGuidedSeg

3 Experiments and Results

Our pelvic dataset comprises 50 prostate cancer patients from a cancer hospital, each accompanied by a T2-weighted MR image and its corresponding manually-annotated label map created by medical experts. Specifically, the prostate, bladder, and rectum in all MRI scans have been manually segmented, serving as ground truth for evaluating our segmentation method. These images were acquired using 3T MRI scanners. The image size is mostly $256 \times 256 \times (120 \sim 176)$, and the voxel size is $1 \times 1 \times 1$ mm^3. A typical example of the MR image and its corresponding label map are given in Fig. 1.

We employ a five-fold cross-validation approach to evaluate our method. In each fold, we randomly select 35 subjects as the training set, 5 subjects as the validation set, and the remaining 10 subjects as the testing set. Unless explicitly mentioned, all reported performance metrics are evaluated on the testing set. For evaluation, we utilize the Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD) to measure the agreement between the manually and automatically segmented label maps.

3.1 Comparison with State-of-the-art Methods

To demonstrate the advantage of our proposed method, we also compare our method with other three widely-used methods on the same dataset as shown in Table 1: 1) SSAE [1], 2) UNet [9], 3) SResSegNet [14].

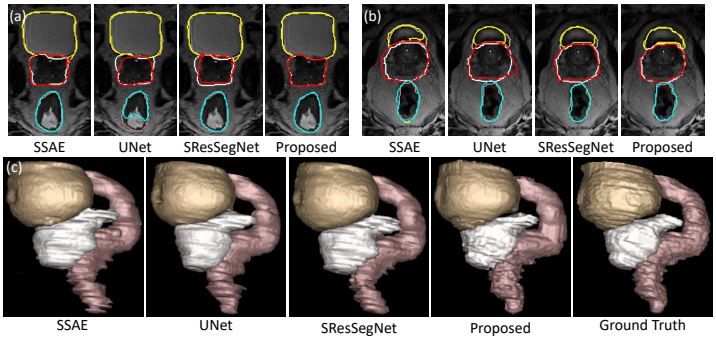

Fig. 4: Visualization of pelvic organ segmentation results by four methods. In (a) and (b), orange, silver and pink contours indicate the manual ground-truth segmentations, while yellow, red and cyan ones indicate automatic segmentations. (a) Clear boundary case, (b) blurry boundary case, and (c) 3D renderings of segmentations.

Table 1: DSC and ASD on the pelvic dataset by four different methods.

Method	DSC			ASD		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
SSAE	.918(.031)	.871(.042)	.863(.044)	1.089(.231)	1.660(.490)	1.701(.412)
UNet	.896(.028)	.822(.059)	.810(.053)	1.214(.216)	1.917(.645)	2.186(0.850)
SResSegNet	.944(.009)	.882(.020)	.869(.032)	.914(.168)	1.586(.358)	1.586(.405)
Proposed	**.975(.006)**	**.932(.017)**	**.918(.025)**	**.850(.148)**	**1.282(.273)**	**1.351(.347)**

Table 1 quantitatively compares our method with three state-of-the-art segmentation methods. We can see that our method achieves better accuracy than the other state-of-the-art methods in terms of both DSC and ASD. It is worth noting that our proposed method can achieve much better performance for the blurry-boundary organ (i.e., prostate), which indicates the effectiveness of our proposed network components for blurry boundary delineation.

We also visualize some typical segmentation results in Fig. 4, which further show the superiority of our proposed method, especially for the blurry regions around the prostate.

3.2 Impact of Each Proposed Component

As our method consists of several novel proposed components, we conduct empirical studies below to analyze them.

Impact of Proposed SG Module: As mentioned in Sec. 2.2, we propose a semantic-guided encoder feature learning module to learn more discriminative features in shallow layers. The effectiveness of the SG module is further confirmed by the improved performance, e.g., 2.40%, 4.41% and 2.8% performance improvements in terms of DSC for bladder, prostate, and rectum, respectively, compared with the UNet with multi-scale deep supervision.

Relationship with Similar Work: Several previous work are proposed to use attention mechanism [7, 10] to enhance the encoder-decoder networks. However, our work is different from them mainly in the follow way: We propose to use highly semantic information from the decoder to explicitly guide the building of attention mechanism, so that we can efficiently learn the encoder features. To further compare them, we visually present the three *typical* learned feature maps (selected by clustering) of a certain layer (i.e., combined layer) in different networks at a certain training iteration (i.e., 4 epochs). The methods include FCN [6], UNet [9], UNet with concurrence SE module [10] (ConSEUNet), attention-UNet [7] (AttUNet) and our proposed one(SGUNet). The visualized maps are in Fig. 5.

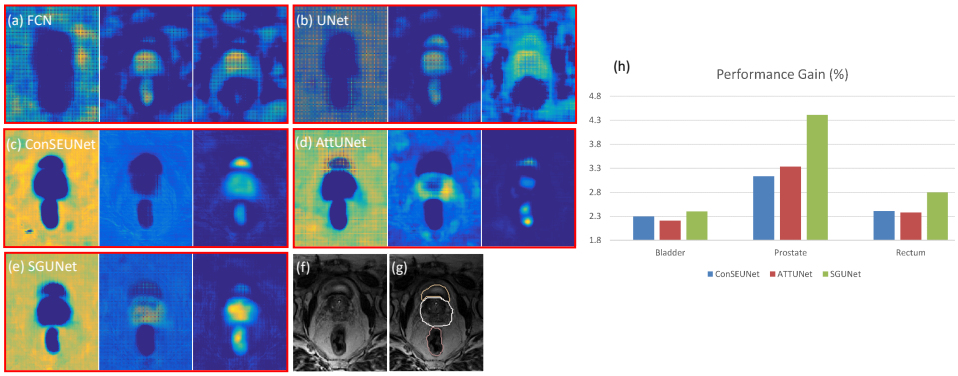

Fig. 5: (a-e): Visualization of three typical learned feature maps of a certain layer by five different networks. (f) and (g) are the corresponding input MRI and the MRI overlaid by the ground-truth contours. (h) is the performance gain in terms of DSC with different strategies towards the UNet with *multi-scale deep supervision*.

Fig. 5(a-e) illustrates that the raw encoder-decoder networks (i.e., FCN and UNet) struggle to handle blurry boundary cases effectively. Attention-based networks produce higher semantic maps with improved localization information. Notably, our proposed method achieves more precise boundaries due to explicit semantic guidance. Additionally, our method demonstrates faster convergence compared to other approaches. Furthermore, the quantitative analysis in Fig. 5(h) is consistent with the the conclusion of qualitative analysis.

Impact of Soft Contour Constraint: As introduced in Sec. 2.3, we apply a hard contour constraint for clear-boundary organs while a soft contour constraint for the blurry-boundary organs. Since hard contour constraint is a widely adopted strategy, we directly compare our proposed soft contour constraint with the case of using hard constraint. With soft constraint on the prostate, we can achieve a

slight performance gain such as 0.2% in terms of DSC; but we can achieve more performance gain in terms of ASD (0.8%), which is mainly because the soft contour constraint can help more accurately locate the blurry boundaries.

3.3 Validation on Extra Dataset

To show the generalization ability of our proposed algorithm, we conduct additional experiments on the PROMISE12-challenge dataset [5]. This dataset contains 50 labeled subjects where only prostate was annotated. We can achieve a high DSC (0.92), small ASD (1.57) in average based on five-fold cross validation. As for the extra 30 subjects' testing dataset whose ground-truth label maps are hidden from us, the performance of our proposed algorithm is still very competitive (we are ranking in the top 6 among 290 submission with an overall score of 89.46. The details can be available via this link[3]) and our solution is compared to the state-of-the-art methods on the 30 subjects' testing dataset [14, 16]. These experimental results indicate a very good generalization capability of our proposed algorithm.

4 Conclusion

In this paper, we have introduced a novel semantic-guided encoder feature learning strategy aimed at capturing highly semantic and rich resolution information features to address the challenge of blurry-boundary delineation. Our SG module enhances the raw skip connection of encoder-decoder models by amplifying discriminative features while attenuating less informative ones. Additionally, we propose a soft contour constraint to model blurry-boundary detection and an ordinary hard contour constraint for clear-boundary detection. This approach has been demonstrated to effectively improve boundary localization and mitigate inter-class errors. By integrating all these proposed components into the network, our final framework has shown significant improvements compared to other methods in terms of both accuracy and robustness, as validated on both the original and extra dataset.

References

[1] Yanrong Guo et al. Deformable mr prostate segmentation via deep feature learning and sparse patch matching. *IEEE TMI*, 35:1077–1089, 2016.
[2] Nicholas Heller et al. Imperfect segmentation labels: How much do they matter? In *MICCAI workshop*, pages 112–120. Springer, 2018.
[3] Victor Lamme et al. Separate processing dynamics for texture elements in primary visual cortex of the macaque monkey. *Cerebral Cortex*, 9(4).
[4] Hong Joo Lee et al. Structure boundary preserving segmentation for medical image with ambiguous boundary. In *CVPR*, pages 4817–4826, 2020.
[5] Geert Litjens et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *MedIA*, 18(2):359–373, 2014.
[6] Jonathan Long et al. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
[7] Ozan Oktay et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
[8] Hariharan Ravishankar et al. Joint deep learning of foreground and shape for robust contextual segmentation. In *IPMI*.
[9] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[3] https://promise12.grand-challenge.org/evaluation/results/

[10] Abhijit Guha Roy et al. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *MICCAI*.

[11] JiCheng Wang et al. Boundary-aware transformers for skin lesion segmentation. In *MICCAI*, pages 206–216. Springer, 2021.

[12] Zelin Wu et al. W-net: A boundary-enhanced segmentation network for stroke lesions. *Expert Systems with Applications*, page 120637, 2023.

[13] Ju Xu, Mengzhang Li, and Zhanxing Zhu. Automatic data augmentation for 3d medical image segmentation. In *MICCAI*, pages 378–387. Springer, 2020.

[14] Lequan Yu et al. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, 2017.

[15] Sihang Zhou et al. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *TIP*, 29:461–475, 2019.

[16] Qikui Zhu et al. Boundary-weighted domain adaptive neural network for prostate mr image segmentation. *arXiv preprint arXiv:1902.08128*, 2019.