FocalUNETR: A Focal Transformer for Boundary-aware Prostate Segmentation using CT Images

Anonymous

No Institute Given

Abstract. Computed Tomography (CT) based precise prostate segmentation for treatment planning is challenging due to (1) the unclear boundary of the prostate derived from CT's poor soft tissue contrast and (2) the limitation of convolutional neural network-based models in capturing long-range global context. Here we propose a novel focal transformer-based image segmentation architecture to effectively and efficiently extract local visual features and global context from CT images. Additionally, we design an auxiliary boundary-induced label regression task coupled with the main prostate segmentation task to address the issue of unclear boundaries from poor soft tissue contrast CT images. We demonstrate that these designs can significantly improve the quality of the CT-based prostate segmentation task over other competing methods, resulting in substantially improved Dice Similarity Coefficient and reduced Hausdorff Distance and Average Symmetric Surface Distance on both private and public CT image datasets.

Keywords: Focal transformer · Prostate segmentation · Computed tomography · Boundary-aware

1 Introduction

Prostate cancer is a leading cause of cancer-related deaths in adult males, as reported in studies, such as [16]. A common treatment option for prostate cancer is external beam radiation therapy (EBRT) [4], where CT scanning is a cost-effective tool for the treatment planning process compared with the more expensive magnetic resonance imaging (MRI). As a result, precise prostate segmentation in CT images becomes a crucial step, as it helps to ensure that the radiation doses are delivered effectively to the tumor tissues while minimizing harm to the surrounding healthy tissues.

Due to the relatively low spatial resolution and soft tissue contrast in CT images compared to MRI images, manual prostate segmentation in CT images can be time-consuming and may result in significant variations between operators. Several automated segmentation methods have been proposed to alleviate these issues, especially the fully convolutional networks (FCN) based U-Net [17] (an encoder-decoder architecture with skip connections to preserve details and

extract local visual features) and its variants [13,21,24]. Despite good progress, these methods often have limitations in capturing long-range relationships and global context information [2] due to the inherent bias of convolutional operations. Researchers naturally turn to ViT [5], powered with self-attention (SA), for more possibilities: TransUNet first [2] adapts ViT to medical image segmentation tasks by connecting several layers of the transformer module (multi-head SA) to the FCN-based encoder for better capturing the global context information from the high-level feature maps. TransFuse [23] and MedT [19] use a combined FCN and Transformer architecture with two branches to capture global dependency and low-level spatial details more effectively. Swin-UNet [1] is the first U-shaped network based purely on more efficient Swin Transformers [11] and outperforms models with FCN-based methods. UNETR [6] and SiwnUNETR [18] are Transformer architectures extended for 3D inputs.

Despite the improved performance for the aforementioned ViT-based networks, these methods utilize the standard or shifted-window-based SA, which is the fine-grained local SA and may overlook the local and global interactions [22]. As reported by [18], even pre-trained with a massive amount of medical data with self-supervised learning, the performance of prostate segmentation task using high-resolution and better soft tissue contrast MRI images is not completely satisfactory, not to mention the lower-quality CT images. Additionally, the unclear boundary of the prostate in CT images coming from the low soft tissue contrast is usually not properly addressed [7,20].

Recently, Focal Transformer [22] is proposed for general computer vision tasks, in which focal self-attention is proposed to incorporate both fine-grained local and coarse-grained global interactions. Each token attends its closest surrounding tokens with fine granularity, and the tokens far away with coarse granularity; thus, focal SA can capture both short- and long-range visual dependencies efficiently and effectively. Inspired by this work, we propose the FocalUNETR (Focal U-NEt TRansformers), a novel focal transformer architecture for CT-based medical image segmentation (Fig. 1A). Despite prior works such as PsiNet [14], which incorporates additional decoders to enhance boundary detection and distance map estimation, these methods still either lack the capacity for effective global context capture through FCN-based techniques or overlook the significance of considering the randomness of the boundary, particularly in poor soft tissue contrast CT images for prostate segmentation. In contrast, our approach utilizes a multi-task learning strategy that leverages a Gaussian kernel over the boundary of the ground truth segmentation mask [10] as an auxiliary boundary-aware contour regression task (Fig. 1B). This serves as a regularization term for the main task of generating the segmentation mask. This auxiliary task enhances the model's generalizability by addressing the challenge of unclear boundaries in low-contrast CT images.

In this paper, we make several contributions. First, we develop a novel focal transformer model (FocalUNETR) for CT-based prostate segmentation, which makes use of focal SA to hierarchically learn the feature maps accounting for both short- and long-range visual dependencies efficiently and effectively. We also

address the challenge of unclear boundaries specific to CT images by incorporating an auxiliary task of contour regression. Our methodology achieves superior performance compared to other state-of-the-art methods, as demonstrated by extensive experiments on both private and public datasets.

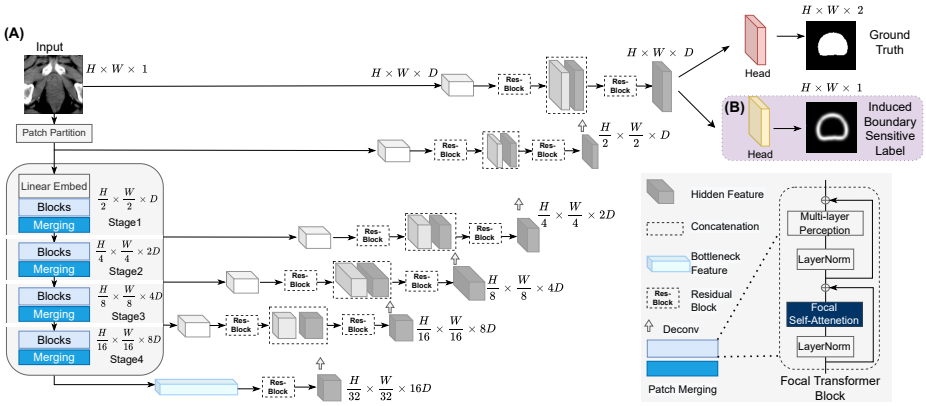

Fig. 1. The architecture of FocalUNETR (A) as the main task for prostate segmentation and a boundary-aware regression auxiliary task (B).

2 Methods

2.1 FocalUNETR

Our FocalUNETR architecture (Fig. 1) follows a multi-scale design similar to [6,18], enabling us to obtain hierarchical feature maps at different stages. The input medical image $\mathcal{X} \in \mathcal{R}^{C \times H \times W}$ is first split into a sequence of tokens with dimension $\lceil \frac{H}{H'} \rceil \times \lceil \frac{W}{W'} \rceil$, where H, W represent spatial height and width, respectively, and C represents the number of channels. These tokens are then projected into an embedding space of dimension D using a patch of resolution (H', W'). The SA is computed at two focal levels [22]: fine-grained and coarse-grained, as illustrated in Fig. 2A. The focal SA attends to fine-grained tokens locally, while summarized tokens are attended to globally (reducing computational cost). We perform focal SA at the window level, where a feature map of $x \in \mathcal{R}^{d \times H'' \times W''}$ with spatial size $H'' \times W''$ and d channels is partitioned into a grid of windows with size $s_w \times s_w$. For each window, we extract its surroundings using focal SA.

For window-wise focal SA [22], there are three terms $\{L, s_w, s_r\}$. Focal levels L is the number of granularity levels for which we extract the tokens for our focal SA. We present an example, depicted in Fig. 2B, that illustrates the use of two focal levels (fine and coarse) for capturing the interaction of local and

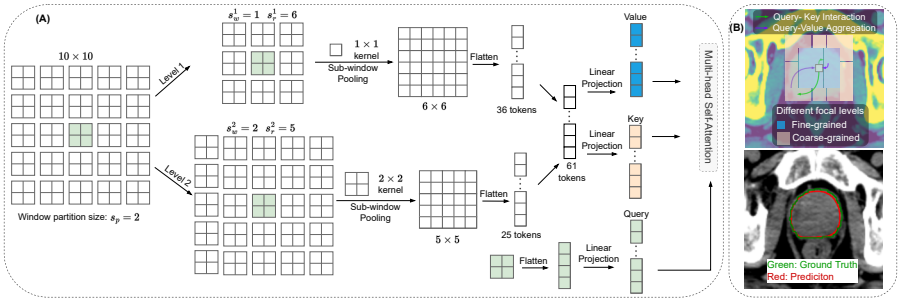

Fig. 2. (A) The focal SA mechanism, and (B) an example of perfect boundary matching using focal SA for CT-based prostate segmentation task (lower panel), in which focal SA performs query-key interactions and query-value aggregations in both fine- and coarse-grained levels (upper panel).

global context for optimal boundary-matching between the prediction and the ground truth for prostate segmentation. Focal window size s_w^l is the size of the sub-window on which we get the summarized tokens at level $l \in \{1, \ldots, L\}$. Focal region size s_r^l is the number of sub-windows horizontally and vertically in attended regions at level l. The focal SA module proceeds in two main steps, sub-window pooling and attention computation. In the sub-window pooling step, an input feature map $x \in \mathcal{R}^{d \times H'' \times W''}$ is split into a grid of sub-windows with size $\{s_w^l, s_w^l\}$, followed by a simple linear layer f_p^l to pool the sub-windows spatially. The pooled feature maps at different levels l provide rich information at both fine-grained and coarse-grained, where $x^l = f_p^l(\hat{x}) \in \mathcal{R}^{d \times \frac{H''}{s_w^l} \times \frac{W''}{s_w^l}}$, and $\hat{x} = \text{Reshape}(x) \in \mathcal{R}^{(d \times \frac{H''}{s_w^l} \times \frac{W''}{s_w^l}) \times (s_w^l \times s_l^w)}$. After obtaining the pooled feature maps ${x^l}_1^L$, we calculate the query at the first level and key and value for all levels using three linear projection layers f_q, f_k and f_v:

$$Q = f_q(x^1), K = \{K^l\}_1^L = f_k(\{x^1, \ldots, x^L\}), V = \{V^l\}_1^L = f_v(\{x^1, \ldots, x^L\}).$$

For the queries inside the i-th window $Q_i \in \mathcal{R}^{d \times s_w \times s_w}$, we extract the $s_r^l \times s_r^l$ keys and values from K^l and V^l around the window where the query lies in and then gather the keys and values from all L to obtain $K_i = \{K_1, \ldots, K_L\} \in \mathcal{R}^{s \times d}$ and $V_i = \{V_1, \ldots, V_L\} \in \mathcal{R}^{s \times d}$, where $s = \sum_{l=1}^{L} (s_r^l)^2$. Finally, a relative position bias is added to compute the focal SA for Q_i by

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + B)V_i,$$

where $B = \{B^l\}_1^L$ is the learnable relative position bias [22].

The encoder utilizes a patch size of 2×2 with a feature dimension of $2 \times 2 \times 1 = 4$ (i.e., a single input channel CT) and a D-dimensional embedding space. The overall architecture of the encoder comprises four stages of focal transformer

blocks, with a patch merging layer applied between each stage to reduce the resolution by a factor of 2. We utilize an FCN-based decoder (Fig. 1A) with skip connections to connect to the encoder at each resolution to construct a "U-shaped" architecture for our CT-based prostate segmentation task. The output of the encoder is concatenated with processed input volume features and fed into a residual block. A final 1×1 convolutional layer with a suitable activation function, such as Softmax, is applied to obtain the required number of class-based probabilities.

For the experiments, we follow the hyperparameter settings suggested in [22], with 2 focal levels, transformer blocks of depths [2, 2, 6, 2], and head numbers [4, 8, 16, 32] for each of the four stages. We then create FocalUNETR-S and FocalUNETR-B with D as 48 and 64, respectively. These settings have parameters of 27.3 M and 48.3 M, which are comparable to other state-of-the-art models.

2.2 The Auxiliary Task

For the main task of mask prediction (as illustrated in Fig. 1A), a combination of Dice loss and Cross-Entropy loss is employed to evaluate the concordance of the predicted mask and the ground truth on a pixel-wise level. The objective function for the segmentation head is given by: $\mathcal{L}_{seg} = \mathcal{L}_{dice}(\hat{p}_i, G) + \mathcal{L}_{ce}(\hat{p}_i, G)$, where $\hat{p}_i$ represents the predicted probabilities from the main task and G represents the ground truth mask, both given an input image i. The predicted probabilities, $\hat{p}_i$, are derived from the main task through the application of the FocalUNETR model to the input CT image.

To address the challenge of unclear boundaries in CT-based prostate segmentation, an auxiliary task is introduced for the purpose of predicting boundary-aware contours to assist the main prostate segmentation task. This auxiliary task is achieved by attaching another convolution head after the extracted feature maps at the final stage (see Fig. 1B). The boundary-aware contour, or the induced boundary-sensitive label, is generated by considering pixels near the boundary of the prostate mask. To do this, the contour points and their surrounding pixels are formulated into a Gaussian distribution using a kernel with a fixed standard deviation of σ (in this specific case, e.g., $\sigma = 1.6$) [12,7,10]. The resulting contour is a heatmap in the form of a $Heatsum$ function [10]. We predict this heatmap with a regression task trained by minimizing mean-squared error instead of treating it as a single-pixel boundary segmentation problem. Given the ground truth of contour G_i^C, induced from the segmentation mask for input image i, and the reconstructed output probability $\hat{p}_i^C$, we use the following loss function: $\mathcal{L}_{reg} = \frac{1}{N}\sum_i ||\hat{p}_i^C - G_i^C||_2$ where N is the total number of images. This auxiliary task is trained concurrently with the main segmentation task.

A multi-task learning approach is adopted to regularize the main segmentation task through the auxiliary boundary prediction task. The overall loss function is a combination of $\mathcal{L}_{seg}$ and $\mathcal{L}_{reg}$: $\mathcal{L}_{tol} = \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_{reg}$, where λ_1 and λ_2 are hyper-parameters that weigh the contribution of the mask prediction

loss and contour regression loss, respectively, to the overall loss. The optimal setting of $\lambda_1 = \lambda_2 = 0.5$ is determined by trying different settings.

3 Experiments and Results

3.1 Datasets and Implementation Details

To evaluate our method, we use a large private dataset with 400 CT scans and a large public dataset with 300 CT scans (AMOS [9]). As far as we know, the AMOS dataset is the only publicly available CT dataset including prostate ground truth. We randomly split the private dataset with 280 scans for training, 40 for validation, and 80 for testing. The AMOS dataset has 200 scans for training and 100 for testing [9]. Although the AMOS dataset includes the prostate class, it mixes the prostate (in males) and the uterus (in females) into a single class labeled PRO/UTE. We filter out CT scans missing the PRO/UTE ground-truth segmentation.

For our implementation, we utilize a server equipped with 8 Nvidia A100 GPUs, each with 40 GB of memory. All experiments are conducted in PyTorch, and each model is trained on a single GPU. We interpolate all CT scans into an isotropic voxel spacing of $[1.0 \times 1.0 \times 1.5]$ mm for both datasets. Houndsfield unit (HU) range of $[-50, 150]$ is used and normalized to $[0, 1]$. Subsequently, each CT scan is cropped to a $128 \times 128 \times 64$ voxel patch around the prostate area, which is used as input for 3D models. For 2D models, we first slice each voxel patch in the axial direction into 64 slices of 128×128 images for training and stack them back for evaluation. For both datasets, we train models for 200 epochs using the AdamW optimizer with an initial learning rate of $5e^{-4}$. An exponential learning rate scheduler with a warmup of 5 epochs is applied to the optimizer. The batch size is set to 24 for 2D models and 1 for 3D models. We use random flip, rotation, and intensity scaling as augmentation transforms with probabilities of 0.1, 0.1, and 0.2, respectively. We report the Dice Similarity Coefficient (DSC, %), 95% percentile Hausdorff Distance (HD, mm), and Average Symmetric Surface Distance (ASSD, mm) metrics.

3.2 Experiments

Comparison with State-of-the-Art Methods. To demonstrate the effectiveness of FocalUNETR, we compare the CT-based prostate segmentation performance with three 2D U-Net-based methods: U-Net[17], UNet++ [24], and Attention U-Net (AttUNet) [15], two 2D transformer-based segmentation methods: TransUNet [2] and Swin-UNet [1], two 3D U-Net-based methods: U-Net (3D) [3] and V-Net [13], and two 3D transformer-based models: UNETR [6] and SiwnUNETR [18]. nnUNet [8] is used for comparison as well. Both 2D and 3D models are included as there is no conclusive evidence for which type is better for this task [20]. All methods (except nnUNet) follow the same settings as FocalUNETR and are trained from scratch. TransUNet and Swin-UNet are the only methods that are pre-trained on ImageNet.

Table 1. Quantitative performance comparison on the private and AMOS datasets with a mean (standard deviation) for 3 runs with different seeds. An asterisk (*) denotes the model is co-trained with the auxiliary contour regression task. The best results with/without the auxiliary task are boldfaced or colored in blue, respectively.

Method	Private			AMOS		
	DSC ↑	HD ↓	ASSD ↓	DSC ↑	HD ↓	ASSD ↓
U-Net	85.22 (1.23)	6.71 (1.03)	2.42 (0.65)	83.42 (2.28)	8.51 (1.56)	2.79 (0.61)
UNet++	85.53 (1.61)	6.52 (1.13)	2.32 (0.58)	83.51 (2.31)	8.47 (1.62)	2.81 (0.57)
AttUNet	85.61 (0.98)	6.57 (0.96)	2.35 (0.72)	83.47 (2.34)	8.43 (1.85)	2.83 (0.59)
TransUNet	85.75 (2.01)	6.43 (1.28)	2.23 (0.67)	81.13 (3.03)	9.32 (1.87)	3.71 (0.79)
Swin-UNet	86.25 (1.69)	6.29 (1.31)	2.15 (0.51)	83.35 (2.46)	8.61 (1.82)	3.20 (0.64)
U-Net (3D)	85.42 (1.34)	6.73 (0.93)	2.36 (0.67)	83.25 (2.37)	8.43 (1.65)	2.86 (0.56)
V-Net (3D)	84.42 (1.21)	6.65 (1.17)	2.46 (0.61)	81.02 (3.11)	9.01 (1.93)	3.76 (0.82)
UNETR (3D)	82.21 (1.35)	7.25 (1.47)	2.64 (0.75)	81.09 (3.02)	8.91 (1.86)	3.62 (0.79)
SwinUNETR (3D)	84.93 (1.26)	6.85 (1.21)	2.48 (0.52)	83.32 (2.23)	8.63 (1.62)	3.21 (0.68)
nnUNet	85.86 (1.31)	6.43 (0.91)	2.09 (0.53)	83.56 (2.25)	8.36 (1.77)	**2.65 (0.61)**
FocalUNETR-S	86.53 (1.65)	5.95 (1.29)	2.13 (0.29)	82.21 (2.67)	8.73 (1.73)	3.46 (0.75)
FocalUNETR-B	87.73 (1.36)	5.61 (1.18)	2.04 (0.23)	83.61 (2.18)	8.32 (1.53)	2.76 (0.69)
FocalUNETR-S*	87.84 (1.32)	5.59 (1.23)	2.12 (0.31)	83.24 (2.52)	8.57 (1.70)	3.04 (0.67)
FocalUNETR-B*	**89.23 (1.16)**	**4.85 (1.05)**	**1.81 (0.21)**	**83.79 (1.97)**	**8.31 (1.45)**	2.71 (0.62)

Quantitative results are presented in Table 1, which show that the proposed FocalUNETR, even without co-training, outperforms other FCN and Transformer baselines (2D and 3D) in both datasets for most metrics on both datasets. When co-trained with the auxiliary contour regression task and using the multi-task training strategy, the performance of FocalUNETRs is further improved. The multi-task training strategy further enhances the performance of FocalUNETRs. In summary, these observations indicate that incorporating FocalUNETR and multi-task training with an auxiliary contour regression task can improve the challenging CT-based prostate segmentation performance.

Qualitative results of several representative methods are visualized in Fig. 3. The figure shows that our FocalUNETR-B and FocalUNETR-B* generate more accurate segmentation results that are more consistent with the ground truth than the results of the baseline models. All methods perform well for relatively easy cases (1^{st} row in Fig. 3), but the FocalUNETRs outperform the other methods. For more challenging cases (rows 2-4 in Fig. 3), such as unclear boundaries and mixed PRO/UTE labels, FocalUNETRs still perform better than other methods. Additionally, the FocalUNETRs are less likely to produce false positives (see more in supplementary materials) for CT images without a foreground ground truth, due to the focal SA mechanism that enables the model to capture global context and helps to identify the correct boundary and shape of the prostate. Overall, the FocalUNETRs demonstrate improved segmentation capabilities while preserving shapes more precisely, making them promising tools for clinical applications.

Ablation Study. To better examine the efficacy of the auxiliary task for FocalUNETR, we selected different settings of λ_1 and λ_2 for the overall loss function $\mathcal{L}_{tol}$ on the private dataset. The results (Table 2) indicate that as the value of λ_2 is gradually increased and that of λ_1 is correspondingly decreased (thereby

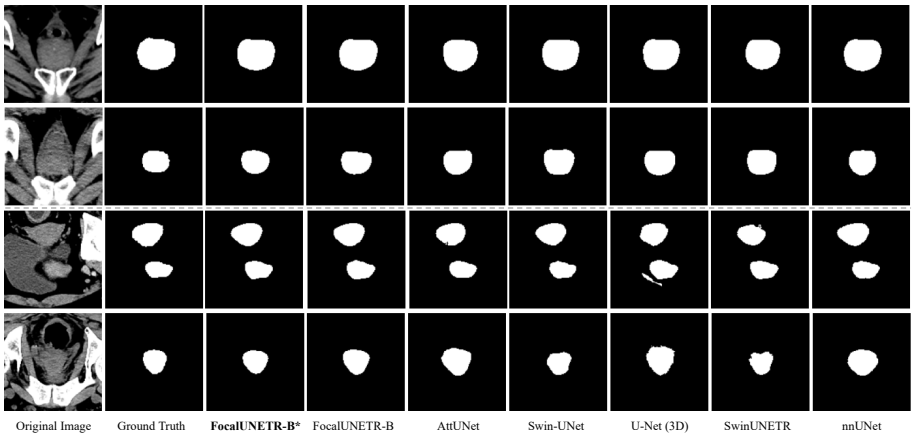

Fig. 3. Qualitative results on sample test CT images from the private (first two rows) and AMOS (last two rows) datasets

increasing the relative importance of the auxiliary contour regression task), segmentation performance initially improves. However, as the ratio of contour information to segmentation mask information becomes too unbalanced, performance begins to decline. Thus, it can be inferred that the optimal setting for these parameters is when both λ_1 and λ_2 are set to 0.5.

Table 2. Ablation study on different settings of total loss for FocalUNETR-B on the private dataset

$\mathcal{L}_{tol}$	$\mathcal{L}_{seg}$	$0.8\mathcal{L}_{seg} + 0.2\mathcal{L}_{reg}$	$0.5\mathcal{L}_{seg} + 0.5\mathcal{L}_{reg}$	$0.2\mathcal{L}_{seg} + 0.8\mathcal{L}_{reg}$
DSC ↑	87.73 ± 1.36	88.01 ± 1.38	**89.23 ± 1.16**	87.53 ± 2.13

4 Conclusion

In summary, the proposed FocalUNETR architecture has demonstrated the ability to effectively capture local visual features and global contexts in CT images by utilizing the focal self-attention mechanism. The auxiliary contour regression task has also been shown to improve the segmentation performance for unclear boundary issues in low-contrast CT images. Extensive experiments on two large CT datasets have shown that the FocalUNETR outperforms state-of-the-art methods for the prostate segmentation task. With regard to this study, limitations need to be noted, including the lack of evaluation for other organs besides the prostate and the need to extend the focal self-attention mechanism for 3D inputs. Future research regarding the 3D version of FocalUNETR for multi-organ segmentation would be worthwhile.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
4. D'Amico, A.V., Whittington, R., Malkowicz, S.B., Schultz, D., Blank, K., Broderick, G.A., Tomaszewski, J.E., Renshaw, A.A., Kaplan, I., Beard, C.J., et al.: Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. Jama **280**(11), 969–974 (1998)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR **abs/2010.11929** (2021)
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
7. He, K., Lian, C., Zhang, B., Zhang, X., Cao, X., Nie, D., Gao, Y., Zhang, J., Shen, D.: Hf-unet: learning hierarchically inter-task relevance in multi-task u-net for accurate prostate segmentation in ct images. IEEE Transactions on Medical Imaging **40**(8), 2118–2128 (2021)
8. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2019)
9. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
10. Lin, L., Wang, Z., Wu, J., Huang, Y., Lyu, J., Cheng, P., Wu, J., Tang, X.: Bsda-net: A boundary shape and distance aware joint learning framework for segmenting and classifying octa images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 65–75. Springer (2021)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
12. Ma, J., Wei, Z., Zhang, Y., Wang, Y., Lv, R., Zhu, C., Gaoxiang, C., Liu, J., Peng, C., Wang, L., et al.: How distance transform maps boost segmentation cnns: an empirical study. In: Medical Imaging with Deep Learning. pp. 479–492. PMLR (2020)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)

14. Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M.: Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In: 2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 7223–7226. IEEE (2019)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. Parikesit, D., Mochtar, C.A., Umbas, R., Hamid, A.R.A.H.: The impact of obesity towards prostate diseases. Prostate international **4**(1), 1–6 (2016)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
19. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)
20. Wang, S., Liu, M., Lian, J., Shen, D.: Boundary coding representation for organ segmentation in prostate cancer radiotherapy. IEEE transactions on medical imaging **40**(1), 310–320 (2020)
21. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME). pp. 327–331. IEEE (2018)
22. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)
23. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2021)
24. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)