SwinFPN: Leveraging Vision Transformers for 3D Organs-At-Risk Detection

Bastian Wittmann[1] BASTIAN.WITTMANN@TUM.DE
Suprosanna Shit[1,3] SUPROSANNA.SHIT@TUM.DE
Fernando Navarro[1,2,3] FERNANDO.NAVARRO@TUM.DE
Jan C. Peeken[2] JAN.PEEKEN@TUM.DE
Stephanie E. Combs[2] STEPHANIE.COMBS@TUM.DE
Bjoern Menze[3] BJOERN.MENZE@UZH.CH

[1] *Department of Informatics, Technical University of Munich, Germany*
[2] *Department of Radio Oncology and Radiation Therapy, Klinikum rechts der Isar, Germany*
[3] *Department of Quantitative Biomedicine, University of Zurich, Switzerland*

Abstract

Current state-of-the-art detection algorithms operating on 2D natural images utilize the relation modeling capability of vision transformers to increase detection performance. However, the feasibility of adapting vision transformers for the 3D medical object detection task remains largely unexplored. To this end, we attempt to leverage vision transformers for organs-at-risk detection and propose a novel feature extraction backbone, dubbed SwinFPN, which exploits the concept of shifted window-based self-attention. We combine SwinFPN with Retina U-Net's head networks and report superior detection performances. Code for SwinFPN will be available in our medical vision transformer library https://github.com/bwittmann/transoar.

Keywords: Vision transformers, Swin transformer, medical object detection.

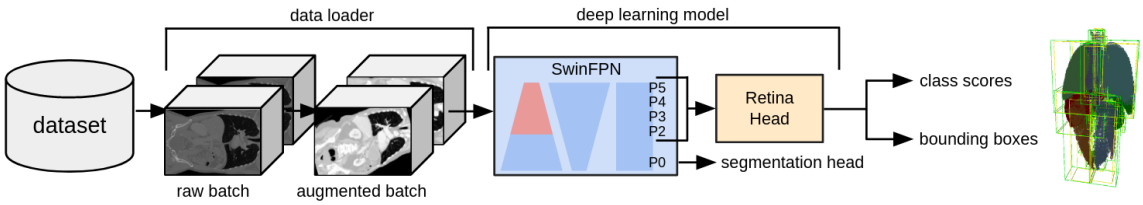

Figure 1: Detection pipeline. A complete overview.

1. Introduction

Radiation therapy represents an effective cancer treatment method that applies high doses of radiation locally to a target volume to damage the DNA of cancer cells beyond repair. However, an extensive patient-specific planning phase is required in which, among other things, organs-at-risk are contoured on CT scans. For this purpose, artificial intelligence-based methods for automated segmentation processing CT scans directly have been proposed. Even though these direct segmentation approaches achieve reasonable results, a preliminary detection step could drastically simplify the segmentation task and hence improve segmentation performances (Navarro et al., 2022).

Recently, nnDetection (Baumgartner et al., 2021), a self-configuring framework for 3D medical object detection, achieved competitive results on prominent benchmarks. The nnDetection framework is built around the one-stage detector Retina U-Net (Jaeger et al., 2020), which is similar to RetinaNet (Lin et al., 2017) and exploits a segmentation proxy task for additional, voxel-wise supervision.
Since the concept of relation modeling should be highly beneficial, especially to capture relative positional dependencies among the regular structured organs-at-risk, we propose a novel vision transformer-based backbone, dubbed SwinFPN, aiming to increase Retina U-Net's detection performance.

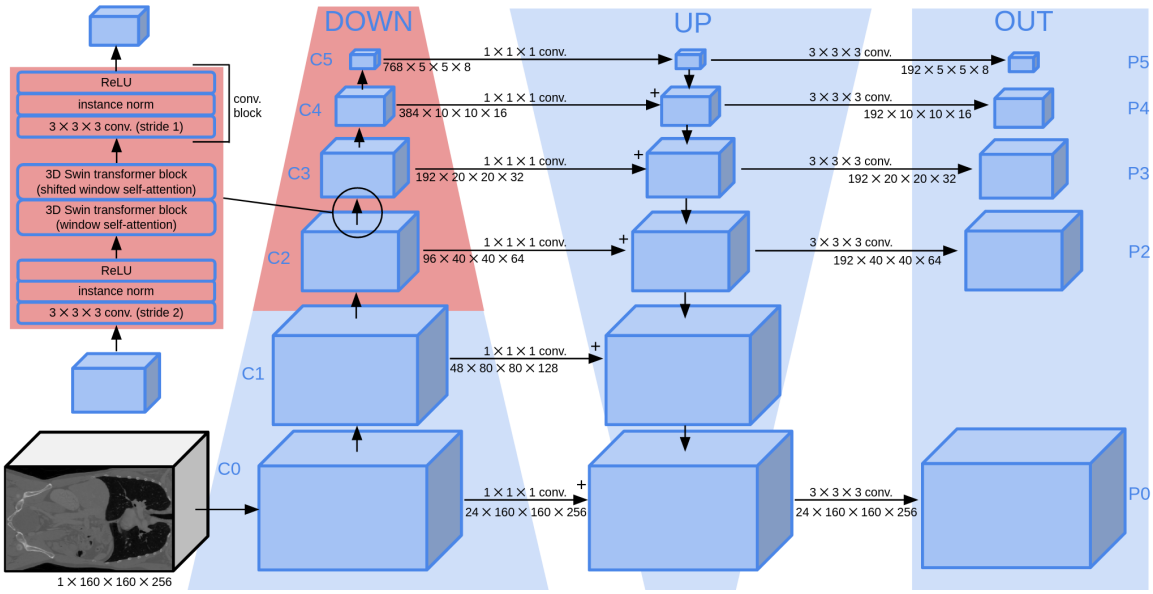

Figure 2: SwinFPN backbone. Regions marked in red utilize 3D Swin transformer blocks.

2. Methodology and Experiments

SwinFPN's detailed architecture is depicted in Figure 2. In general, SwinFPN is represented by a lightweight feature pyramid network consisting of a down-sampling branch (C0 - C5), an up-sampling branch, and a final output projection (P0 - P5). Lateral connections between the down- and up-sampling branches adjust the number of channels of the multi-level feature maps used for detection and classification to the desired value $C_{out} = 192$. After the feature map C1, the standard down-sampling block consisting of two convolutional blocks is augmented with two consecutive 3D Swin transformer (Liu et al., 2021) blocks. Both 3D Swin transformer blocks, which are placed between the two convolutional blocks, exploit the window-based self-attention mechanism with a window size of $7 \times 7 \times 7$. In addition, the windows are shifted by $(3, 3, 3)$ voxels in the second 3D Swin transformer block to enable cross-window information exchange. The feature maps P5, P4, P3, P2, and P0 are finally forwarded to the head networks, as displayed in Figure 1.
The dataset used for the task of 3D organs-at-risk detection consists of CT scans from the VISCERAL anatomy benchmark (Jimenez-del Toro et al., 2016). During pre-processing,

all CT scans were cropped to the thorax and abdomen region and resized to the shape $160 \times 160 \times 256$.
AP and mAP results of experiments with different backbone configurations combined with Retina U-Net's head networks are listed in Table 1. We compare SwinFPN to a standard FPN by simply deactivating the additional 3D Swin transformer blocks. Since SwinFPN possesses a larger amount of trainable parameters, we conduct additional experiments to ensure that the increased detection performance is not just a result of increased model capacity. To this end, we adjust the standard FPN by increasing C_{out} from 192 to 384 (FPN-384) and by adding additional convolutional blocks (FPN-192*).
One can conclude that SwinFPN performs superior compared to the different standard FPN configurations, even though its increase in parameters remains the lowest.

Table 1: Quantitative results. Comparing SwinFPN to standard FPN configurations.

Backbone	Params	$\Delta \downarrow$	$\text{mAP}_{\text{coco}} \uparrow$	$\text{mAP}^{\text{S}}_{\text{coco}} \uparrow$	$\text{mAP}^{\text{M}}_{\text{coco}} \uparrow$	$\text{mAP}^{\text{L}}_{\text{coco}} \uparrow$	$\text{AP}_{50} \uparrow$	$\text{AP}_{75} \uparrow$
FPN-192	44.1M	–	40.64	18.91	**40.31**	59.22	80.67	37.19
FPN-384	68.5M	+24.4M	41.06	**22.17**	39.33	59.41	80.71	37.13
FPN-192*	65.2M	+21.1M	41.43	20.40	40.08	60.98	80.45	36.29
SwinFPN-192	63.1M	**+19.0M**	**41.83**	21.00	39.96	**62.00**	**82.94**	**40.11**

* standard down-sampling block is augmented with an additional conv. block after C1.

3. Conclusion

This work lays the foundation for vision transformer-based 3D medical object detectors. It demonstrates that it is possible to leverage the global relation modeling mechanism for the task of medical object detection with minor adjustments. Although we present initial, promising results, large-scale annotated CT datasets or tailor-made unsupervised pre-training strategies would be necessary to completely exploit the potential of vision transformers.

References

Michael Baumgartner et al. nndetection: A self-configuring method for medical object detection. In *MICCAI*, 2021.

Paul F Jaeger et al. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *ML4H*, 2020.

Oscar Jimenez-del Toro et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. In *IEEE*, 2016.

Tsung-Yi Lin et al. Focal loss for dense object detection. In *ICCV*, 2017.

Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, 2021.

Fernando Navarro et al. A unified 3d framework for organs at risk localization and segmentation for radiation therapy planning. *CoRR, abs/2203.00624*, 2022.