MT-WilmsNet: A Multi-Level Transformer Fusion Network for Wilms' Tumor Segmentation and Metastasis Prediction

Zhu Zhu[1] *, Wenjing Yu[2] *, Xiaohui Ma[1], Shuai Liu[2], Jie Dong[2], Yuxin Du[2], Changmiao Wang[3] **, and Gang Yu[1] **

[1] Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2] Hangzhou Dianzi University, Hangzhou, China
[3] Shenzhen Research Institute of Big Data, Shenzhen, China

yugbme@zju.edu.cn, cmwangalbert@gmail.com

Abstract. Wilms' tumor (WT) is a prevalent cancer affecting the kidneys of children, and accurate segmentation and prediction of metastasis are vital for treatment planning and prognosis. Current methods for assessing metastasis, such as invasive biopsies and expensive PET-CT scans, hinder their widespread use in clinical settings. Deep learning, especially classification models for 3D data, is currently widely used in tumor metastasis prediction. However, existing models may not have fully accounted for the global significance of cross-sectional slices, and segment-assisted classification frameworks tailored for low-cost clinical CT imaging protocols remain understudied, with systematic validation in clinical settings yet to be comprehensively established. In this study, we propose MT-WilmsNet, a slice-guided multi-task multi-level Transformer fusion network featuring three synergistic components. First, a Wide Reinforced Transformer Feature Pyramid Network integrates multi-scale features to boost preoperative metastasis prediction accuracy. Second, a dedicated UNet-like architecture performs tumor segmentation while providing anatomical context for metastasis analysis. Finally, a global slice attention mechanism combined with multi-level self-distilling transformers emulates radiologists' cross-slice diagnostic reasoning. Our MT-WilmsNet outperforms many typical classification models for WT metastasis prediction. The source code is available at: https://github.com/wenjing-gg/MT-WilmsNet.

Keywords: Wilms' Tumor · Multi-task Model· Metastasis Prediction · Multi-level Fusion.

1 Introduction

Wilms' tumor, the most common malignant kidney tumor in children, is clinically diagnosed primarily through imaging examinations, including abdominal plain

* Co-first authors.
** Corresponding author.

radiography, excretory urography, abdominal ultrasound, and abdominal CT or MRI scans [6]. Among these, abdominal non-contrast and contrast-enhanced CT scans constitute the most critical diagnostic procedures. A crucial step in WT management is the assessment of metastasis, which currently relies on invasive biopsies and costly PET-CT imaging [18]. However, these limitations underscore the need for noninvasive, cost-effective alternatives. Techniques based on CT imaging offer significant potential for improving preoperative metastasis risk assessment, thereby optimizing treatment plans and enabling more personalized surgical strategies.

Radiomics develops machine learning models by extracting radiographic features from the CT scans. However, this methodology may lack the capacity to fully harness the deep visual-semantic information inherent in CT scans, leading to limitations in metastatic evaluation for WT. CT image processing using deep learning typically follows one of three approaches: The 2D processing paradigm [21,2,16], which analyzes images slice by slice, benefits from computational efficiency but fails to capture spatial correlations between layers; The 3D processing paradigm [24,20,4], which handles volumetric data in three dimensions, this approach retains complete spatial information but struggles with high memory usage and computational complexity, lacking effective multi-tasking solutions such as classification and segmentation for WT diagnosis and treatment as well. The 2.5D processing paradigm [5,23,10] creates pseudo-3D inputs by combining multiple adjacent slices. Yet, it lacks theoretical guidance on key parameters, like the number of neighboring slices, and is ineffective at modeling cross-slice spatial relationships [10].

To address these challenges, this paper proposes a multi-level self-distilling Transformer fusion network for Wilms' tumor segmentation and metastasis prediction. It integrates global information with focused attention on local slices while exploring multi-level complementary features simultaneously. The main contributions of this work are summarized as follows: (1) A Global Slice Attention (GSA) module with mixed positional encoding is designed for dynamic inter-slice relevance quantification. (2) We propose a Wide Reinforced Transform Feature Pyramid Network (WRT-FPN) for hierarchical feature fusion, which enables adaptive fusion of cross-scale semantic information while preserving spatial resolution. (3) An UNet-like encoder-decoder is introduced for WT segmentation, which captures WT spatial structures and leverages multi-level guidance effectively.

2 Methodology

The proposed MT-WilmsNet model, as shown in Fig. 1, consists of four main components: the GSA module, a Hierarchical 3D-Vision Transformer backbone, a UNet-like encoder-decoder, and the WRT-FPN module. The GSA module integrates information across slices using high-level semantic features, mimicking physicians' attention. The Hierarchical 3D-Vision Transformer extracts discriminative features through multi-head self-attention. The UNet-like encoder-

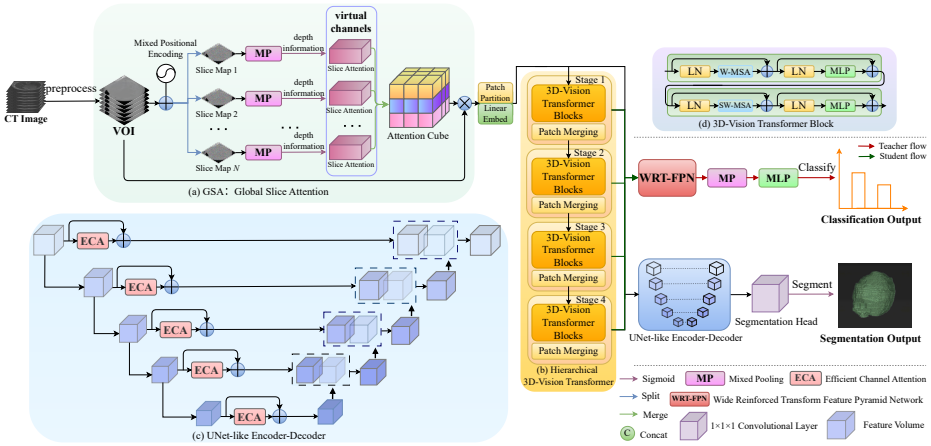


Fig. 1. Overview of the MT-WilmsNet architecture. The student flow extracts feature maps at each level, processes them through a feedforward network, and generates classification outputs for auxiliary supervision. The teacher flow produces the final predictions using WRT-FPN while distilling knowledge into the student stream to enhance its learning capability. The WRT-FPN module will be introduced in Section 2.4.

decoder captures the tumor's spatial structure for segmentation, while the WRT-FPN module fuses multi-level features to address morphological differences, ensuring accurate WT metastasis prediction.

2.1 Global Slice Attention

The GSA module is an attention mechanism designed for preprocessed the volume of interest (VOI) regions. Its structure is illustrated in Fig. 1(a). The module's core idea is to treat the number of channels and slices in a single-channel 3D image as virtual channels, enabling depth information extraction while dynamically weighting each depth slice. This approach highlights critical regions, preserves the original 3D tensor structure, and captures global contextual information. Implementation involves a mixed positional encoding scheme that synergistically integrates depth-aware and spatial-aware encoding components, which are additively fused with the original image tensor. Global information along the depth dimension is extracted through dual streams of global average pooling and maximum pooling. After reshaping, a one-dimensional convolution models dependencies within the depth dimension, followed by point-by-point weighting of the original image.

2.2 Hierarchical 3D-Vision Transformer

The backbone of the proposed model is the Hierarchical 3D-Vision Transformer, as depicted in Fig. 1(b). The input to the Hierarchical 3D-Vision Transformer

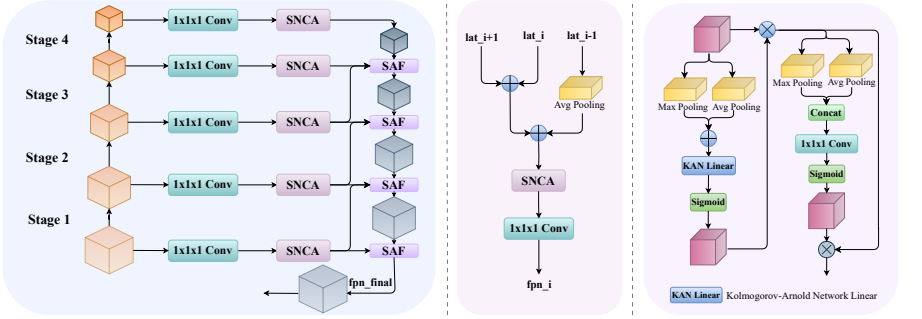
encoder is the 3D feature map processed by the GSA module. This encoder consists of four stages, with each stage containing two 3D-Vision Transformer Blocks. These blocks leverage the Window-based Multi-head Self-Attention (W-MSA) and Shifted Window-based Multi-head Self-Attention (SW-MSA) modules to compute attention weights within the specified area. The shift operation is efficiently computed using a 3D Periodic Shift [13]. In the first stage, a linear embedding layer generates 3D tokens with a resolution of $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. To maintain the hierarchical structure, each stage concludes with a patch-merging layer, which reduces the resolution of the feature representation by half. Specifically, the patch-merging layer groups patches of $2 \times 2 \times 2$ resolution, producing a $4C$ dimensional feature embedding, where C is the number of feature channels. A subsequent linear layer then reduces the feature size from $4C$ to $2C$. Stages 2, 3, and 4 follow the same design, progressively downsampling the resolution to $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$, and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$, enhancing the representation of multi-level features.

2.3 UNet-like Encoder-Decoder

The task of segmenting WT serves as an agent to enhance the model's ability to understand the spatial structure of WT. This task is crucial because the involvement of neighboring organs, local lymph node metastases, or distant metastases reflects the tumor's spatial shape and structural information. As depicted in Fig. 1(c), the decoder component first doubles the resolution of the encoder's output feature maps at each level using transposed convolutional layers. These outputs are then residually concatenated with those from the previous stage. The concatenated features are processed through another residual block, as previously described. Finally, a $1 \times 1 \times 1$ convolutional layer was used as the segmentation head to produce the final segmentation output, followed by a Sigmoid activation function.

2.4 Wide Reinforced Transform Feature Pyramid Network

WT invasion and metastasis patterns can vary significantly across different spatial slices. To address the multi-scale challenges inherent in WT datasets, this study proposed WRT-FPN. As shown in Fig. 2, the WRT-FPN module comprises feature maps at various levels filtered and combined within the feature selection module. Here, the Spatial Non-linear Convolutional Attention (SNCA) module initiates the processing of the encoded feature map $f_{\text{in}} \in \mathbb{R}^{C \times D \times H \times W}$, where D is the data depth, H is the height, and W is the width. The WRT-FPN enhances the spatial relationships within the feature map, intensively weights the feature channels, and further improves spatial non-linear representation capability through the introduced KAN [14] Linear layer. Subsequently, these enhanced feature maps, containing high-level and low-level information from the current and neighboring stage encoders, are processed through the Spatial Adaptive Fusion (SAF) module for local feature path fusion. Finally, the classification results



(a) WRT-FPN: Wide Reinforced Transform Feature Pyramid Network (b) SAF: Spatial Adaptive Fusion (c) SNCA:Spatial Non-linear Convolutional Attention

Fig. 2. Overview of the WRT-FPN module, where SAF is a feature fusion module that receives inputs from the current feature value lat_i and neighboring feature values $lat\_i \pm 1$.

are derived through horizontal convolution, global pooling, and other operations on the feature maps from the final stage.

2.5 Joint Loss Function

To effectively balance the loss across multiple subtasks during training, this study utilizes an uncertainty-weighted composite loss function [12]. This approach is mathematically represented in Equation (1):

$$\mathcal{L}_{joint} = \sum_{i=1}^{N} \left[\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right], \tag{1}$$

where each loss component $\mathcal{L}_i$ comprises $\mathcal{L}_{cls}$, $\mathcal{L}_{seg}$ and $\mathcal{L}_{aux}$. A higher σ_i indicates more significant uncertainty and difficulty in the subtask, reducing its relative weight in total loss.

Classification Loss: To improve generalization and reduce overfitting, especially with small datasets, this study employs Label Smoothing cross-entropy loss. This technique smooths the distribution of categories within the dataset, with a smoothing coefficient set at 0.1.

Segmentation Loss: This study combines voxel-level cross-entropy loss with Dice loss based on region overlap to address complex shapes and class imbalance challenges, as shown in Equation (2):

$$\begin{aligned} \mathcal{L}_{seg}(\mathbf{z}, \mathbf{y}) = & \underbrace{\alpha \left[-\frac{1}{NV} \sum_{n=1}^{N} \sum_{i=1}^{V} \log(\sigma(\mathbf{z}_n)_{y_{n,i}}) \right]}_{\mathcal{L}_{CE}} \\ & + \underbrace{\beta \left[1 - \frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{2 \sum_{i=1}^{V} p_{n,c,i} t_{n,c,i}}{\sum_{i=1}^{V} p_{n,c,i} + \sum_{i=1}^{V} t_{n,c,i} + \gamma} \right]}_{\mathcal{L}_{Dice}}, \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ denotes the softmax activation function, N represents the number of samples, V denotes the number of voxels, $\mathcal{L}_{CE}$ represents the standard cross-entropy loss and $\mathcal{L}_{Dice}$ encapsulates the Dice similarity coefficient in a 3D context. In this study, both α and β are set to 1.

Auxiliary Loss: The auxiliary loss accelerates training convergence through self-distillation. Specifically, the composite loss function consists of two components: (a) Student Flow: The first term is the accumulated label-smoothed cross-entropy loss between predictions from hierarchical encoder layers and ground-truth labels. (b) Teacher Flow Alignment: The second term employs Kullback-Leibler (KL) divergence regularization, ensuring that the multi-layer predictions align with the probabilistic distribution of the final output, which the teacher model guides.

3 Experiments and Results

3.1 Dataset and Implementation

The Wilms' tumor dataset. The scarcity of publicly available datasets hinders current WT imaging studies. For instance, existing databases like SEER [11] provide only clinical feature data without imaging information, while mainstream medical imaging datasets, such as KiTS23 [8], focus on renal tumors but lack detailed annotation granularity. These datasets typically distinguish between tumors and cysts but do not include critical labels for metastatic status, which are essential for accurate analysis. To address these limitations, we retrospectively assembled a multi-center, annotated Wilms' tumor CT dataset comprising 197 postoperative pediatric cases imaged between January 2012 and December 2024. All patients underwent contrast-enhanced abdominal CT before any surgery, biopsy, radiotherapy, or chemotherapy. Among the initial cohort, 109 cases were metastatic and 86 non-metastatic. We excluded studies with missing or low-quality scans (e.g., motion artifacts), preoperative treatment, or ambiguous diagnoses. All data were anonymized for analysis. Of these, 80% were randomly allocated for model training and 20% for testing. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Academic Ethics Committee of Children's Hospital Zhejiang University School of Medicine (IRB No. 2023-IRB-0287-P-01; granted 16 Nov 2023). We have applied for an informed consent waiver for our study.

Data preprocessing. A phased VOI extraction strategy was used. In training, lesion segmentation labels guided automatic boundary extraction, expanded by 20% to retain peri-tumoral information before standardization. During inference, physicians manually outlined 3D bounding boxes based on WT imaging, followed by standardization: (1) intensity normalization within [-100, 200] Hounsfield Unit and (2) resampling to 64^3 voxels via spatial interpolation. This strategy ensures spatial consistency and standardized model inputs.

Implementation details. All experiments were conducted using PyTorch 2.4.0 on an NVIDIA RTX 4090 GPU with 24 GB of memory. Our model utilized

pre-trained weights from SwinUNETR [7], with the backbone of the original SwinViT network frozen during training. Training proceeded for 200 epochs, optimized using AdamW with a weight decay of 1e-4 to prevent overfitting. We employed a customized learning rate scheduler based on the WarmupCosineLR strategy. The learning rate increased linearly from 1e-5 to 2e-5 during the first 10 epochs (warmup phase), then decayed to 0 via cosine annealing for the remaining epochs. The batch size was set to 2, but each sample underwent five augmentation techniques—including random rotation, Gaussian noise addition, and intensity scaling—effectively diversifying the training data and simulating a larger batch size of 10.

3.2 Experiment Analysis

Quantitative Comparison. The proposed model's performance is evaluated in classification and segmentation through comparative analyses. First, we compared it with traditional radiomics methods, followed by several 3D-CNN-based networks. Finally, we benchmarked it against state-of-the-art SAM-based models. These models include MedicalNet [3], SwinUNETR [7], SAM-Med3D [19], MAPSeg [22], VIVIT [1] and MTS-Net [9]. To ensure fairness, we used open-source code for comparison methods with default parameters and matched data volume to training cycles. The classification was evaluated using AUC, accuracy, specificity, sensitivity, and F1-score, while segmentation was assessed with DSC, JI, ASD, and HD95, as shown in Table 1. Our method demonstrated superior performance across most metrics compared to other models. While traditional radiomics methods performed poorly with multicenter data, VIVIT models without transfer learning showed the lowest classification metrics. Notably, our approach achieved a 13% improvement in AUC for classification, surpassing mainstream 3D image classification models. Additionally, it showed a relative increase of at least 16% in the comprehensive F1 metric, achieving segmentation performance on par with leading models and surpassing competing methods in overall accuracy. These results underscore the significant advancements our approach offers in multi-task performance.

Ablation Study. To assess the effectiveness of each component within MT-WilmsNet, we conducted experiments by training a base model and incrementally adding combinations of the GSA module, multi-task architecture, and the WRT-FPN module. The results are presented in Table 2. Our analysis of various modules uncovered several key insights. The GSA module enhances the model's ability to capture the significance of each slice, boosting AUC by 4% over the baseline. The multi-task framework enhances tumor localization, aiding metastasis prediction. Notably, the WRT-FPN module significantly strengthens feature fusion, yielding a 6% AUC gain for WT metastasis predictions, ultimately optimizing overall performance.

Table 1. Quantitative comparison with different models on the private WT dataset. The optimal results are shown in bold, and the sub-optimal results are underlined.

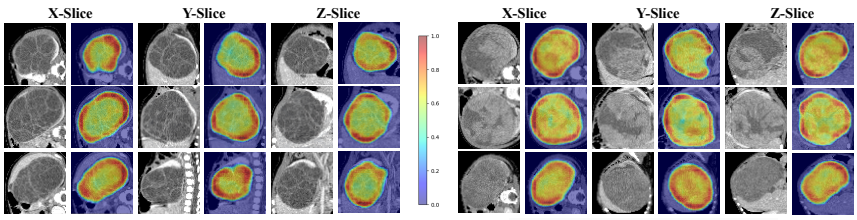
Model	AUC ↑	ACC ↑	Specificity ↑	Sensitivity ↑	F1-score ↑	DSC ↑	JI↑	ASD ↓	HD95 ↓
Classification									
VIVIT [1]	0.5432	0.5411	0.6444	0.3225	0.3855	\	\	\	\
Radiomics [17]	0.6414	0.5752	0.5101	0.6364	0.6222	\	\	\	\
MedicalNet [3]	<u>0.7313</u>	0.7013	<u>0.7652</u>	0.4477	0.5541	\	\	\	\
MTS-Net [9]	0.6907	<u>0.7250</u>	0.7222	<u>0.7273</u>	<u>0.7442</u>	\	\	\	\
Segmentation									
MAPSeg [22]	\	\	\	\	\	0.8543	0.7491	4.2043	27.7468
SwinUNETR [7]	\	\	\	\	\	0.8861	0.8104	3.3934	14.6179
SAM-Med3D [19]	\	\	\	\	\	<u>0.9225</u>	<u>0.8574</u>	<u>0.6931</u>	<u>2.9094</u>
MT-WilmsNet (Ours)	**0.8712**	**0.8501**	**0.7778**	**0.9091**	**0.8696**	**0.9231**	**0.8597**	**0.6452**	**2.7188**

Table 2. Ablation study performance with progressively added modules.

w/o	AUC ↑	ACC ↑	Specificity ↑	Sensitivity ↑	F1 Score ↑	DSC ↑	JI ↑	ASD ↓	HD95 ↓
Baseline	0.7551	0.7250	0.8333	0.6364	0.7179	\	\	\	\
+GSA	0.7904	0.7500	0.6111	0.8636	0.7917	\	\	\	\
+Multi-task	0.8157	0.8496	**0.8889**	0.8182	0.8571	0.9211	0.8566	0.6699	2.7753
+WRT-FPN	**0.8712**	**0.8501**	0.7778	**0.9091**	**0.8696**	**0.9231**	**0.8597**	**0.6452**	**2.7188**

3.3 Thermal Map Visualization

In this section, we utilized 3D-GradCam [15] to visualize the focus regions during the model's evaluation of WT metastasis. Fig. 3(a) and Fig. 3(b) illustrate the visualization results for cases without and with metastasis, respectively. The visualization demonstrates that our model effectively concentrates on the tumor's edges and accurately identifies the tumor's location. This focused attention at the tumor's boundaries allows the model to locate anomalies and meticulously predict metastasis.



(a) Cases of Wilms' tumor without metastasis (b) Cases of Wilms' tumor with metastasis

Fig. 3. Visualization of original images and heatmaps. Each row represents the difference between the 3D data in different orientations, containing the original voi image and the corresponding attention heat map.

4 Conclusion

In this study, a multi-level Transformer fusion network for Wilms' tumor segmentation and metastasis prediction named MT-WilmsNet is introduced. It integrates the Global Slice Attention (GSA) module for dynamic inter-slice relevance quantification and the Wide Reinforced Transform Feature Pyramid Network (WRT-FPN) for adaptive cross-scale semantic fusion. Additionally, a UNet-like Encoder-Decoder leverages multi-level guidance and WT segmentation as an auxiliary task to capture spatial structure. Comparative and ablation experiments on an internal dataset demonstrate the model's superiority and effectiveness in WT metastasis prediction and segmentation. With further validation, we aim to implement this approach as a supportive tool for staging prediction in WT.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (No. 2023YFC2706400), the Guangxi Key R&D Project (No. AB24010167), the Project (No. 20232ABC03A25), Zhejiang Provincial Natural Science Foundation (No. LTGY23F020003), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617).

Disclosure of Interests. The authors declare no competing interests.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
2. Bhattarai, B., Subedi, R., Gaire, R.R., Vazquez, E., Stoyanov, D.: Histogram of Oriented Gradients meet deep learning: A novel multi-task deep network for 2D surgical image semantic segmentation. Medical Image Analysis **85**, 102747 (2023)
3. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
4. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. Medical Image Analysis **98**, 103324 (2024)
5. Guo, X., Chen, M., Zhou, L., Zhu, L., Liu, S., Zheng, L., Chen, Y., Li, Q., Xia, S., Lu, C., et al.: Predicting early recurrence in locally advanced gastric cancer after gastrectomy using CT-based deep learning model: a multicenter study. International Journal of Surgery **111**(2), 2089–2100 (2025)
6. Han, Q., Li, K., Dong, K., Xiao, X., Yao, W., Liu, G.: Clinical features, treatment, and outcomes of bilateral Wilms' tumor: A systematic review and meta-analysis. Journal of Pediatric Surgery **53**(12), 2465–2469 (2018)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)

8. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Medical Image Analysis **67**, 101821 (2021)
9. Huang, Y., Jin, Y., Tao, K., Xia, K., Gu, J., Yu, L., Du, L., Chen, C.: MTS-Net: Dual-enhanced positional multi-head self-attention for 3D CT diagnosis of May-Thurner syndrome. arXiv preprint arXiv:2406.04680 (2024)
10. Hung, A.L.Y., Zheng, H., Zhao, K., Du, X., Pang, K., Miao, Q., Raman, S.S., Terzopoulos, D., Sung, K.: CSAM: A 2.5D cross-slice attention module for anisotropic volumetric medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5923–5932 (2024)
11. Institute, N.C.: Surveillance, epidemiology, and end results (SEER) program. Cancer Statistics, SEER Data & Software, Registry Operations (2018)
12. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7482–7491 (2018)
13. 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)
14. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: KAN: Kolmogorov-Arnold Networks. arXiv preprint arXiv:2404.19756 (2024)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**, 336–359 (2020)
16. Senkyire, I.B., Gedeon, K.K., Freeman, E., Ghansah, B., Liu, Z.: EcD-Net: Encoder-Corollary Atrous Spatial Pyramid Pooling-decoder network for automated pancreas segmentation of 2D CT images. Informatics in Medicine Unlocked **51**, 101597 (2024)
17. Tafuri, B., De Blasi, R., Nigro, S., Logroscino, G.: Explainable machine learning radiomics model for Primary Progressive Aphasia classification. Frontiers in Systems Neuroscience **18**, 1324437 (2024)
18. Vujanić, G.M., Sandstedt, B.: The pathology of Wilms' tumour (nephroblastoma): the International Society of Paediatric Oncology approach. Journal of Clinical Pathology **63**(2), 102–109 (2010)
19. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision. pp. 51–67. Springer (2025)
20. Yu, K., Sun, L., Chen, J., Reynolds, M., Chaudhary, T., Batmanghelich, K.: DrasCLR: A self-supervised framework of learning disease-related and anatomy-specific representation for 3D lung CT images. Medical Image Analysis **92**, 103062 (2024)
21. Yue, X., Huang, X., Xu, Z., Chen, Y., Xu, C.: Involving logical clinical knowledge into deep neural networks to improve bladder tumor segmentation. Medical Image Analysis **95**, 103189 (2024)
22. Zhang, X., Wu, Y., Angelini, E., Li, A., Guo, J., Rasmussen, J.M., O'Connor, T.G., Wadhwa, P.D., Jackowski, A.P., Li, H., et al.: MAPSeg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3D masked autoencoding and pseudo-labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5851–5862 (2024)

23. Zhang, Y., Liao, Q., Ding, L., Zhang, J.: Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5D solutions. Computerized Medical Imaging and Graphics **99**, 102088 (2022)
24. Zhang, Z., Keles, E., Durak, G., Taktak, Y., Susladkar, O., Gorade, V., Jha, D., Ormeci, A.C., Medetalibeyoglu, A., Yao, L., et al.: Large-scale multi-center CT and MRI segmentation of pancreas with deep learning. Medical Image Analysis **99**, 103382 (2025)