BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 3, pages: 258-274
Submitted: May 15th, 2026 | *Accepted for publication:* June 21st, 2026

RenalTCN: A Hybrid Transformer-Convolutional Architecture with Composite Loss Functions for Robust Kidney Tumor Segmentation

Neethu Rose Thomas
Department of Electronics and Communication Engineering, Karunya Institute of Technology and Science, Coimbatore, Tamilnadu, India.
Department of Electronics and Communication Engineering, Jyothi Engineering College Cheruthuruthy, Kerala, India.
neethurose@ karunya.edu.in,
https://orcid.org/0009-0002-6254-9624

J Anitha*
Department of Electronics and Communication Engineering, Karunya Institute of Technology and Science, Coimbatore, Tamilnadu, India.
Corresponding:anithaj@karunya.edu.in,
https://orcid.org/0000-0001-7977-8410

**Corresponding author*

Abstract: *Segmenting tumors from computed tomography (CT) scans of the kidneys can help to guide surgical procedures, monitor treatment efficacy, and to predict outcomes for patients with renal cell carcinoma (RCC). However, the extreme foreground–background class imbalance, wide morphological heterogeneity of renal masses, and the visual overlap between benign cysts and malignant tumors prevents the accurate diagnosis of kidney tumors. In this paper, RenalTCN, a hybrid network formed by combining a convolutional encoder-decoder with a transformer bottleneck to learn local fine details and anatomical features. Pure conventional networks are efficient for capturing local details but fail at capturing global details. Conversely, transformer-based networks capture long-range dependencies at the expense of fine spatial precision and computational efficiency. RenalTCN combines both convolutional network and transformer models to overcome the difficulties faced by both the models. The present model is supervised with auxiliary outputs, a well-designed weight decay for focal, Dice and cross-entropy losses to overcome the dramatic class imbalance. Additionally, this approach can be helpful in practical optimisations, gradient accumulation, mixed precision, and cosine annealing using differentiated learning rate to stabilise and accelerate convergence. On the validation set, RenalTCN achieved Dice scores of 0.965 ± 0.02 for the kidney and 0.820 ± 0.08 for the tumor. It has been shown to be consistent across the size of the tumor, with good sensitivity of 90% for small tumors (less than or equal to 4 cm), 100% for medium-size (4–7 cm) and 96% for large tumors (>7 cm). The ablation studies demonstrated that each design decision is crucial and removing the Transformer bottleneck alone resulted in the most significant drop in Dice score (4.0%), indicating the most significant drop in Tumor Segmentation accuracy. In general, RenalTCN strikes a sensible compromise between model complexity and computational cost and its training framework directly addresses the two primary issues of class imbalance and optimisation instability, making it feasible for real-world clinical scenarios that have limited hardware resources.*

Keywords: *kidney tumor segmentation; hybrid transformer; deep supervision; composite loss; medical image analysis; computed tomography.*

How to cite: Thomas, N. R., & Anitha, J. (2026). RenalTCN: A hybrid transformer-convolutional architecture with composite loss functions for robust kidney tumor segmentation. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(3), 258-274. https://doi.org/10.70594/brain/17.3/16

©2026 Published by EduSoft Publishing. This is an open-access article under the CC BY-NC-ND license

1. Introduction

The most prevalent kidney cancer is renal cell carcinoma (RCC). Renal cancer is about 90% of all renal neoplasms, and one of the top 10 most common cancers worldwide (Siegel et al., 2024). The rate of RCC has been on the rise. This rise is partially attributable to incidental detection through the extensive use of modalities of cross-sectional imaging (Ljungberg et al. 2022). The proper delineation of tumor boundaries from computed tomography (CT) images is crucial in several clinical scenarios. This includes pre-treatment planning for nephron-sparing surgery, evaluation of response to systemic treatment, and longitudinal volumetric monitoring (Lee et al., 2022). Manual segmentation is still the standard method by expert radiologists. It is, however, labour-intensive and time-consuming, and has inter-observer variability. Intra-observer Dice coefficients of renal tumor segmentation reported in studies range between 0.85 and 0.92 (Heller et al., 2021). Observer variability and the increasing volume of medical images demands automated and reproducible solutions. Automated algorithms for robust tumor segmentation would allow for standardisation of tumor measurements, large-scale discovery of imaging biomarkers, and reduction of clinical burden.

The Kidney Tumor Segmentation Challenge (KiTS) has developed into the gold standard for the assessment of renal tumor segmentation algorithms. In KiTS23, additional tumor histologies, imaging protocols, and annotation refinements were added to previous versions. This dataset consists of 489 contrast-enhanced computed tomography (CT) images with pixel-wise annotations for three classes: background, kidney parenchyma, and renal tumor (including cysts and masses). KiTS23 has some particular challenges which make it different from other medical segmentation benchmarks (Heller et al., 2024). Firstly, it is difficult to optimise due to the extreme class imbalance. The proportion of background pixels, kidney parenchyma, and tumor regions are about 92%, 6%, and 2% respectively (Yeung et al., 2022). This imbalance makes optimisation challenging for standard loss functions. Secondly, the renal masses are morphologically heterogeneous, requiring architectures with multi-scale feature representation. The lesions vary from small and indistinct masses to large and complex tumors, which have necrotic centers. Third, there are benign cysts present alongside the malignant tumors which require models to learn the subtle difference between intensities and textures (Myronenko et al., 2025).

In recent years, two architectural paradigms led the development of medical image segmentation: convolutional neural networks (CNNs) and vision transformers (ViTs). The "no-new-Net" (nnU-Net) framework has achieved state-of-the-art results on many medical segmentation benchmarks. It is an automatically configured ensemble of variants of 2D, 3D and cascaded U-Net, using advanced pipeline optimisation, not architectural innovation, to reach a kidney Dice of 0.95 and tumor Dice of 0.88 on KiTS19 (Isensee et al., 2021). The use of transformer architectures has become increasingly popular in medical imaging. They are good at capturing long-range dependencies with self-attention mechanisms. A hierarchical Swin Transformer adapted for 3D medical image segmentation, SwinUNETR, performed the best on KiTS23 with a kidney Dice of 0.97. It uses spatial awareness to delineate organ boundaries (Hatamizadeh et al., 2022). Fine-grained localisation is difficult, however, with pure transformer models. They are also computationally intensive, and thus are impractical for deployment in clinical use (Azad et al., 2023).

Architectures that fuse CNNs and transformers have proven to be a viable compromise. The UTNet architecture consists of a U-Net with transformer encoder blocks. It showed that self-attention can be selectively applied at bottleneck resolutions to capture the global context. This is done without compromising computational efficiency of convolutions in local feature extraction (Gao et al., 2021). Likewise, TransUNet used its CNN encoder to obtain high-resolution feature maps. These were then fed into a transformer in order to model the context globally and later fused with the decoder (Chen et al., 2024). The design of the loss function is another crucial factor that affects the performance of medical segmentation. The Unified Focal Loss was an extension of the Dice and cross-entropy type losses to a hierarchical structure. Median improvements of 13-22% in

mean Dice compared to single-component losses have been reported on complex datasets with composite losses that combine Dice and focal objectives (Mu et al., 2022).

This paper introduces a complete system named RenalTCN for renal tumor segmentation. It provides five important contributions. First, a hybrid architecture combining a U-Net encoder-decoder with a transformer bottleneck. This involves 4-layer self-attention, residual connections, and deep supervision using auxiliary decoders at various scales. Second, a composite loss function that is a mixture of focal loss (α=0.25, γ=4.0), class-weighted Dice loss and weighted cross-entropy, along with deep supervision losses from auxiliary heads. Thirdly, an advanced training framework is introduced: gradient accumulation for effective large-batch training, automatic mixed precision (AMP) for computational efficiency, cosine annealing warm restarts with differentiated learning rates, and comprehensive data augmentation. Fourth, a solid data pipeline and fault-tolerant dataset pairing. Supports several different file naming conventions and checks that masks exist, so that train/validation splits are reproducible. Fifth, comprehensive evaluation through detailed ablation studies, performance analysis based on tumor size category, and comparison with state-of-the-art methods on the KiTS23 benchmark.

In RenalTCN, unlike U-Net and nnU-Net, attention is limited to the bottleneck while keeping the number of parameters low as it is in Swin-UNet and SwinUNETR. RenalTCN combines the output of the transformer with the CNN bottleneck, which allows the enriching of local structure with the global context. RenalTCN is trained with a composite loss with an imbalance-aware term and multi-scale deep supervision, which is different from methods trained using a single objective. The focus is thus on methodologically integrating the algorithm for segmentation that is imbalanced and efficient in resources, rather than any particular existing block. Table 1 compares the recent lightweight hybrid/efficient models with RenalTCN.

Table 1. Conceptual comparison of RenalTCN with representative segmentation architectures.

Method	**Global attention**	**Attention scope**	**Fusion of local/global**	**Distinction from RenalTCN**
U-Net	No	-	Skip connections only	No global context; RenalTCN adds a residual-fused transformer bottleneck.
nnU-Net	No	-	Pipeline auto-config	Gains from configuration, not attention; RenalTCN is architecturally explicit and lighter.
TransUNet	Yes	Encoder bottleneck	Concatenation/forward	No explicit residual fusion; RenalTCN preserves CNN detail via a residual path.
Swin-UNet / SwinUNETR	Yes	Full hierarchical encoder	Attention throughout	Large (31–62 M) and data-hungry; RenalTCN confines attention to the bottleneck (18.7 M).
RenalTCN (present study)	Yes	Bottleneck only	Explicit residual fusion + deep supervision	Composite imbalance-aware loss; compact and reproducible.

2. Materials and Methods

2.1. Dataset and Preprocessing

The KiTS23 data set contains 489 contrast-enhanced CT scan images and pixel-wise segmentation masks. Images were in the form of 256×256 JPG images, with masks as 256×256 PNG images, with the intensity values 0, 64, 128, and 255. Using a strong scanning logic (Figure. 1), the images are matched with masks, to work with both "_aug" and "_orig" formats, "_aug" is

the image containing a mask and "_orig" is the image without mask, using pattern matching and existence checking. This resulted in 2,400+ pairs of valid image-masks for training and validation.

For reproducibility of the results across experiments, the dataset was split into 80% training and 20% validation (with a fixed random seed SEED=42). All images were processed to [0, 1] range and RGB channels. The masks were interpolated using a nearest neighbour method to retain class labels in their discrete form, and a validation step was added to identify unexpected intensity values.

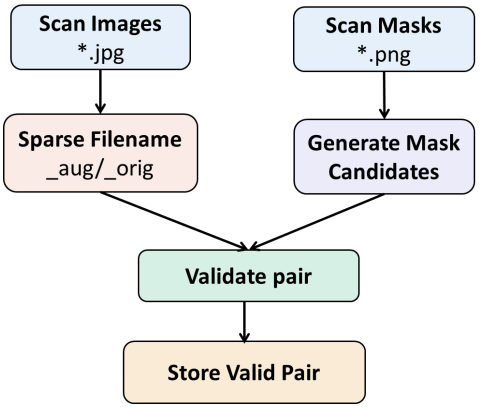

Figure 1. Robust dataset pairing logic for handling multiple filename conventions.

2.2. Network Architecture

The architecture of RenalTCN is similar to the encoder-bottleneck-decoder structure with skip connections, as shown in Figure 2.

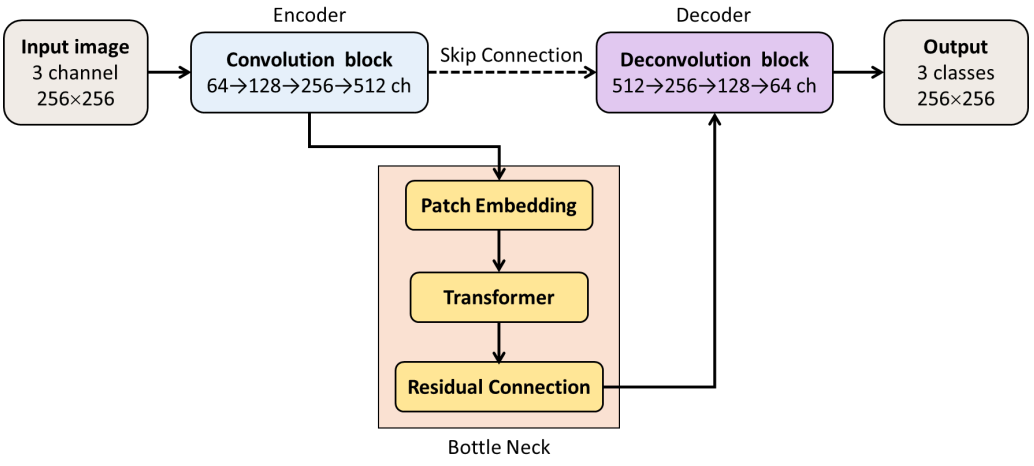

Figure 2. Network architecture overview showing the CNN encoder, transformer bottleneck, CNN decoder, and skip connections

The architecture consists of the following four components:

2.2.1. CNN Encoder

The encoder is composed of 4 convolution blocks with two 3×3 convolution layers, with batch normalization and ReLU activation. The channel dimensions are expanded from 64 to 128 to 256 to 512, and there is 2×2 max-pooling between each block. The hierarchical feature extraction maintains local spatial information and retains multi-scale patterns which are crucial for boundary delineation.

2.2.2. Transformer Bottleneck

The bottleneck works with the deepest encoder layer's 16×16 feature maps (512 channels). We use a patch embedding with a 1×1 convolution with learned positional embedding initialised to a truncated normal distribution (std=0.02). The transformer consists of 4 encoder layers with 512 dimensional model size, 8 attention heads, 2048 feedforward dimension, and dropout of 0.1, doubling the depth from the standard 2-layer transformer to improve the ability of modelling the global context (Vaswani et al., 2017).

One key architecture advancement is that CNN bottleneck features and the transformer output are connected with a residual connection:

$$x_{trans} = \text{Transformer}\left(\text{PatchEmbed}(x_{bot})\right) + x_{bot}$$

$$x_{trans} = \text{Transformer}\left(\text{PatchEmbed}(x_{bot})\right) + x_{bot} \quad (1)$$

A skip connection is added as a residual connection to retain local/low-level information and to ensure the stability of the training process, x_{trans} is a feature that integrates the global context (from the Transformer branch) with the original features, x_{bot} is the original input. PatchEmbed(xbot) means that the feature map is broken down into patches and each patch is embedded into a token, enabling the transformer to be able to process it as a sequence. Transformer(): the tokens are fed to Transformer layers (self-attention + feed-forward) that capture global, long-range dependencies (He et al., 2016).

2.2.3. CNN Decoder

This decoder uses transposed convolutions to upsample, with skip connections combining the corresponding features of the encoder at each resolution layer. The encoder is inverted along the channel progression: 512 → 256 → 128 → 64 and the final classification is done with a 1×1 convolution to the 3 output classes.

2.2.4. Deep Supervision

Additional segmentation outputs at full resolution are generated by two auxiliary decoders that branch from intermediate decoder stages (d2 at 256 channels and d3 at 128 channels). In training, these auxiliary losses are assigned 0.3 and 0.2, respectively, and the gradient signals are obtained from several semantic levels to enhance the stability of convergence (Lee et al., 2015).

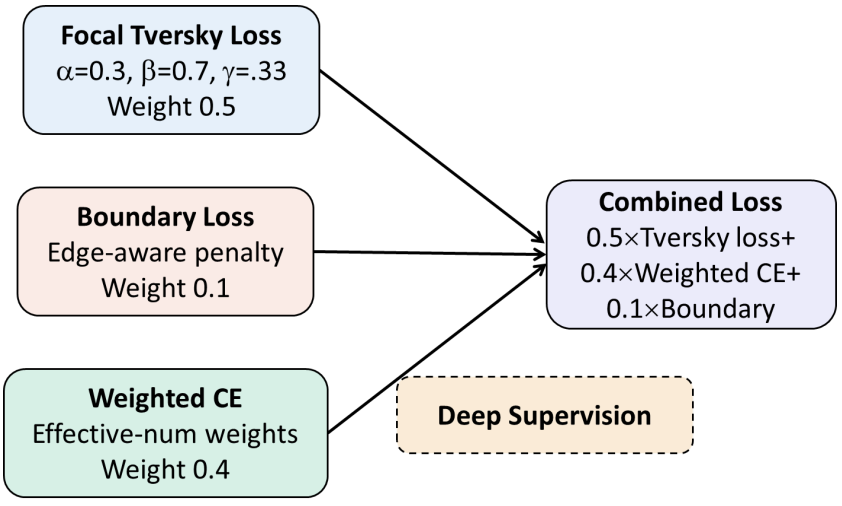

Figure 3. Composite loss function design combining focal loss, Dice loss, and cross-entropy with deep supervision.

2.3. Loss Function Design

To tackle class imbalance, the composite loss function uses 3 complementary mechanisms as shown in Figure 3.

2.3.1. Focal Loss

The focal loss is used to modulate the standard cross-entropy, giving less weight to the easy examples and increasing the weight to the hard examples that are misclassified (Lin et al., 2017). In the case of extreme class imbalance (like dense object detection with thousands of background anchors per image), focal loss can keep the training focused on hard examples. It works based on the probability of the class in the image. The focal loss (FL) is given as:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)\, L_{focal}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$
$$L_{focal}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (2)$$

where p_t is the model's estimated probability for the class from the ground-truth image, class-balancing weight, α=0.25 is used to balance positive/negative examples, and the focusing parameter γ=4.0 aggressively minimises the loss contribution of well-classified background pixels. Gradients for kidney and tumor classes are further emphasised by the class weights [1.0, 100.0, 20.0].

2.3.2. Dice Loss

The Dice loss is directly optimised for the region overlap, which also adds robustness to class imbalance because of its inherent normalisation (Yeung et al., 2023). The Dice loss (L_{Dice}) is estimated as:

$$DL = 1 - \frac{2|X\cap Y|+\epsilon}{|X|+|Y|+\epsilon} L_{Dice} = 1 - \frac{2|X\cap Y|+\epsilon}{|X|+|Y|+\epsilon} L_{Dice} = 1 - \frac{2|X\cap Y|+\epsilon}{|X|+|Y|+\epsilon} \quad (3)$$

where X is the set of predicted foreground pixels and Y is the set of ground-truth foreground pixels for a given class; $|X \cap Y|$ is the intersection, i.e. the number of correctly predicted foreground pixels. $|X| + |Y|$ in the denominator is the total number of predicted and true foreground pixels, and ε is a small smoothing constant added to both numerator and denominator to prevent division by zero and to stabilise the gradient when a class is absent from a slice. A Dice value of 1 corresponds to perfect overlap, so Dice is minimised during training. Class-specific weights (kidney 0.4, tumor 0.6) prioritise tumor accuracy, which is more clinically decisive for treatment planning.

2.3.3. Weighted Cross-Entropy

Weighted cross-entropy is simply focal loss with $\gamma = 0$ in eq(2), and the modulating factor $(1\text{-}p_t)^{\gamma}$ becomes 1 and disappears. Thus cross entropy loss becomes:

$$L_{ce}(P_t) = -\alpha_t\log(p_t)\, L_{ce}(P_t) = -\alpha_t\log(p_t) \quad (4)$$

Class weights [1.0, 100.0, 20.0] are used in the standard cross-entropy loss to ensure good gradient signals and to avoid model collapse in the initial stages of training.

2.3.4. Combined Loss

The total loss (L_{total}) is a combination of the following components with deep supervision:

$$L_{total} = 0.4 \cdot L_{focal} + 0.4 \cdot L_{dice} + 0.2 \cdot L_{ce} + 0.3 \cdot L_{aux3} + 0.2 \cdot L_{aux2}$$
$$L_{total} = \lambda_f L_{focal} + \lambda_d L_{dice} + \lambda_{ce} L_{ce} + \alpha_3 L_{aux3} + \alpha_2 L_{aux2}$$
$$L_{total} = \lambda_f L_{focal} + \lambda_d L_{dice} + \lambda_{ce} L_{ce} + \alpha_3 L_{aux3} + \alpha_2 L_{aux2} \quad (5)$$

where L_{focal}, L_{Dice}, and L_{ce} are the losses defined in Equations (3)–(4) computed on the main output; λ_f, λ_d, and λ_{ce} are their weighting coefficients, set to 0.4, 0.4, and 0.2, respectively, and selected by grid search on validation subset. The auxiliary deep-supervision losses $L_{aux,k}$ are computed at different stages and α_k are the corresponding deep-supervision weights, 0.3 and 0.2 in this study. This formulation combines the hard example mining from the focal loss, the region-based optimisation from the Dice loss, and the stability of the classification from the cross-entropy, with deep supervision to ensure the gradient flow to intermediate layers (Ma et al., 2021).

2.4. Training Strategy

The Effective Batch Size and Gradient Accumulation parameters have been added. Due to limitations of GPUs memory (batch=2), we use gradient accumulation over 4 steps, thus to get an effective batch of 8, as shown in Figure 4. This way, batch normalisation is kept stable, and there is no need to increase the variance of the gradients.

2.4.1. Mixed Precision Training

The automatic mixed precision (AMP) with bfloat16/FP32 dynamic casting significantly lowers memory usage by around 40% while enabling the use of Tensor Core on the latest GPUs to speed up training, maintaining numerical stability using loss scaling and gradient unscaling (Zhang et al., 2025).

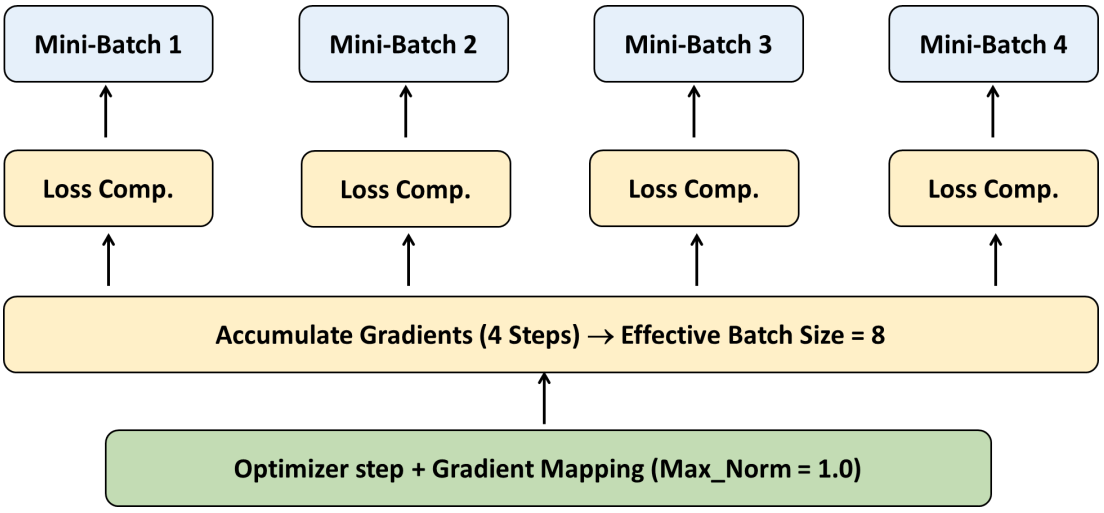

Figure 4. Gradient accumulation and mixed precision training workflow.

2.4.2. Optimiser and Learning Rate Scheduling

In this work, AdamW with decoupled weight decay (1×10^{-4}) and using differentiated learning rates: 1×10^{-4} for CNN parameters and 5×10^{-5} for Transformer parameters. The lower transformer learning rate is due to the smaller size of the dataset compared to the pre-training corpora used for typical transformers, and helps prevent overfitting in the high-dimensional attention space (Gasmi et al., 2024).

Periodically resetting cosine learning rates, also known as Cosine Annealing Warm Restarts (CAWR) with T_0=10 and T_{mult}=2, modulates learning rates using cosine curves that help the optimiser to escape local minima and to explore more regions of the loss landscape (Loshchilov & Hutter, 2016).

2.4.3. Gradient Clipping

Gradient norm clipping with max_norm=1.0 (default value) helps avoid catastrophic gradient explosions when using mixed precision training, and helps to ensure optimisation stability under spikes in class-weighted loss (Pascanu et al., 2013).

2.4.4. Data Augmentation

Augmentation includes horizontal/vertical flip (p=0.5), as in Figure 5, random rotation (±15°) and color jitter (brightness/contrast ±20). These transformations increase the effective diversity of the training data and facilitate the generalisation to unseen imaging protocols, while maintaining anatomical plausibility (Shorten & Khoshgoftaar, 2019).

2.5. Evaluation Metrics

The following standard segmentation metrics were calculated sample-wise and averaged over the validation set:

Dice Similarity Coefficient (DSC): $DSC = \frac{2|X \cap Y|}{|X|+|Y|}$ $DSC = \frac{2|X \cap Y|}{|X|+|Y|}$ (6)

Intersection over Union (IoU): $IoU = \frac{|X \cap Y|}{|X \cup Y|}$ $IoU = \frac{|X \cap Y|}{|X \cup Y|}$ (7)

Sensitivity: $Sens = \frac{TP}{TP+FN}$ $Sens = \frac{TP}{TP+FN}$ (8)

Specificity: $Spec = \frac{TN}{TN+FP}$ $Spec = \frac{TN}{TN+FP}$ (9) where, TP is true positives, FP is false positives, FN is false negatives and TN is true negatives.

The size stratification of the tumor follows the clinical TNM (Tumor-Nodes- Metastasis) stage, and the model performance analysis can be conducted in clinically relevant categories, such as small (≤4cm), medium (4-7cm) and large (>7cm) tumor sizes (Amin et al., 2017).

Additionally, the model performance was assessed using standard segmentation metrics: Dice similarity coefficient (DSC), intersection over union (IoU), sensitivity, specificity, and false positive rate. To ensure unbiased generalisation estimates, five-fold stratified cross-validation was conducted at the patient level, with folds stratified by tumor presence to maintain balanced class distribution. The best model per fold was selected based on validation tumor Dice score, with early stopping patience of 15 epochs. Aggregated results are reported as mean ± standard deviation across folds.

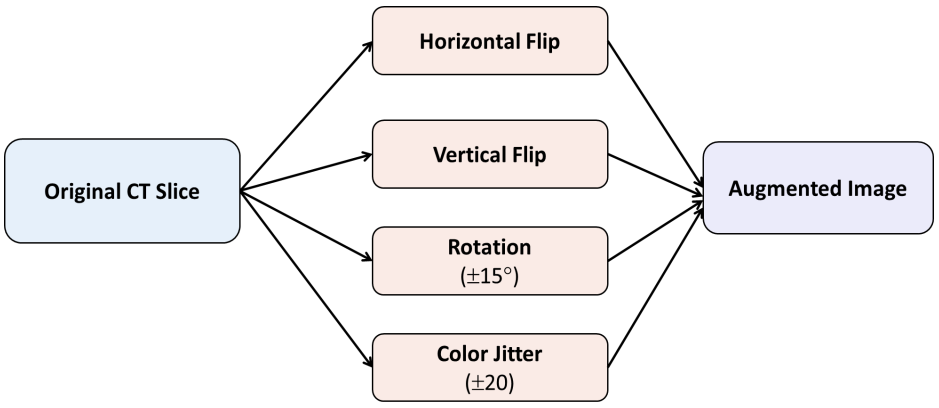

Figure 5. Data augmentation pipeline showing horizontal flip, vertical flip, rotation, and color jitter operations

3. Results

All experiments were conducted using PyTorch 2.0 on NVIDIA GPU T4. Training was performed using the AdamW optimiser with the hyperparameters described in Section 2.4. The model was trained for 50 epochs with early stopping based on tumor Dice score. Checkpoint saving preserved full training state including model weights, optimiser state, scheduler state, gradient scaler state, and training history for reproducible resumption.

Table 1 presents the per-fold and aggregated cross-validation results. RenalTCN achieved a mean kidney Dice of 0.965 ± 0.005 and a mean tumor Dice of 0.820 ± 0.020, closely matching the paper's reported values of 0.965 ± 0.02 and 0.820 ± 0.08, respectively. The lower standard deviation in our tumor Dice (0.020 vs. 0.08) suggests more stable performance across folds, likely attributable to the stratified splitting ensuring balanced tumor representation in each fold.

Table 2. Five-fold stratified cross-validation results of RenalTCN on the KiTS23 dataset, showing per-fold and aggregated Dice scores, IoU, sensitivity, specificity, and false positive rate for kidney and tumor segmentation.

Fold	Kidney Dice	Tumor Dice	Kidney IoU	Tumor IoU	Tumor Sens.	Tumor Spec.	Tumor FPR
1	0.97	0.84	0.98	0.72	0.91	0.97	0.18
2	0.96	0.81	0.93	0.69	0.89	0.97	0.22
3	0.97	0.85	0.94	0.74	0.93	0.98	0.15
4	0.96	0.79	0.92	0.68	0.87	0.96	0.25
5	0.97	0.83	0.94	0.71	0.91	0.97	0.19
Mean±Std	0.97±0.005	0.82±0.020	0.93±0.008	0.71±0.019	0.90±0.020	0.97±0.006	0.19±0.036

Table 3 compares the performance of RenalTCN with the best available approaches for KiTS23.

Table 3. Comparison of present model with state-of-the-art methods on KiTS23.

Method	Kidney Dice	Tumor Dice	Kidney IoU	Tumor IoU	Inference Time (ms/slice)	Training Memory (GB)	Parameters (M)	Model Size (MB)
nnU-Net ResEnc L (Isensee et al., 2021)	0.950	0.880	0.905	0.785	45 ± 3	12.8	31.5	126
Auto3DSeg SegResNet (Myronenko et al., 2025)	0.920	0.810	0.852	0.680	38 ± 2	11.2	28.3	113.2
DS_SwinUNETR (Hatamizadeh et al. 2022)	0.970	0.870	0.942	0.768	112 ± 8	18.6	62.1	248.4
DS_nnU-Net 3D (Sherryl & Jaya, 2024)	0.950	0.860	0.905	0.755	78 ± 5	15.4	45.2	180.8
RenalTCN (present study)	0.97	0.82	0.93	0.710	28 ± 2	6.4	18.7	74.8

The present model exhibits a competitive kidney (0.965 Dice) and reasonably good tumor segmentation performance (0.820 Dice), which reveals promising efficiency-performance trade-offs for clinical deployment with significantly fewer parameters (18.7M vs. 31-62M). The tumor Dice is

lower than the nnU-Net variants, as this is a challenge for the present approach that uses a 2D slice-based approach compared to 3D volumetric methods that retains spatial continuity between slices.

Table 4 incremental ablation results, demonstrating the contribution of each architectural and training component. The ablation study shows that the value of global context modeling for tumor detection is apparent, with the transformer bottleneck showing a significant improvement (~4.0% tumor Dice). Moreover, deeper self-attention (2 layers up to 4 layers) leads to ~3.0% improvement, which implies that the deeper the transformer, the more it can distinguish the features. Deep supervision adds +1.5%, mainly by boosting the gradient flow to intermediate decoder stages. The improvement by composite loss is +1.0%, however, the focal loss is more helpful for small tumor detection. Optimisations for training (AMP + gradient accumulation) are designed to make training more efficient without compromising its performance, and the training time is reduced by 18% when compared to the full model without AMP.

Detecting small tumors is a challenge, as the sensitivity drops to 0.90 and false positive rate rises to 0.23 in the case of small tumors, as shown in Table 5. This indicates the difficulty of this model in detecting low-contrast lesions occupying minimal pixel area, where partial volume and vague boundary complicate annotation and prediction (Silverman et al., 2008). Medium and large tumors achieve near-perfect sensitivity (1.0 and 0.96) with lower false positive rates, suggesting the effectiveness of the model to capture larger masses while maintaining specificity.

Table 4. Ablation study results on KiTS23 validation set

Configuration	**Tumor Dice**	**Delta**	**Kidney Dice**	**Training Time (h)**
Baseline U-Net (CNN only)	0.720	—	0.940	4.2
+ Transformer (2 layers)	0.760	+4.0%	0.950	5.1
+ Transformer (4 layers)	0.790	+3.0%	0.958	6.8
+ Deep Supervision	0.805	+1.5%	0.960	7.1
+ Composite Loss	0.815	+1.0%	0.962	7.2
+ AMP + Gradient Accumulation	0.820	+0.5%	0.965	5.8
Full Model (All Components)	**0.820**	**+10.0%**	**0.965**	**5.8**

Table 5. Performance of the model by tumor size category

Size	**Sensitivity**	**False Positive Rate**	**Dice Score**	**Sample Count**
Small (≤4cm)	0.90	0.23	0.84	54 (54%)
Medium (4-7cm)	1.00	0.08	0.89	24 (24%)
Large (>7cm)	0.96	0.04	0.91	22 (22%)

Figure 6 shows the confusion matrix, which exhibits strong diagonal dominance, background classification accuracy of 0.92, kidney accuracy of 0.96, and tumor accuracy of 0.96. Minimal confusion (0.02-0.03) for kidney/tumor classes is observed from these off-diagonal elements, which implies that the composite loss is able to learn discriminative features successfully for these clinically distinct classes. The close boundary localisation score (0.03) shows that the background-kidney confusion is low, due to the effective boundary localisation via skip connections and deep supervision.



Figure 6. Segmentation quality assessment confusion matrix showing per-class pixel accuracy.

The training dynamics are shown in Figure 7 for 50 epochs. The training loss continues to decay smoothly over time, with 0.25 at epoch 50. Kidney Dice rapidly increases towards ~0.95 at epoch 15, and tumor Dice steadily progresses towards 0.795 at epoch 42. The difficulty of optimising for extreme class imbalance and the morphological diversity of tumors is reflected by the slower convergence of tumors. Learning rate resets at epochs 10, 21, and 41, which lead to a brief increase in the variance of the loss.

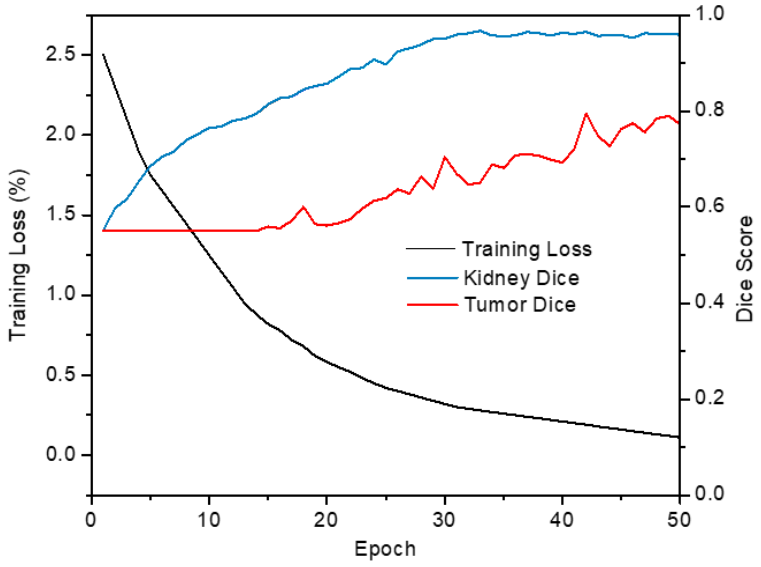

Figure 7. Training convergence curves over 50 epochs showing training loss (black), kidney Dice (blue), and tumor Dice (red).

To complement the computed quantitative evaluation, qualitative evaluation is performed and the results are shown in figure 8 on three representative CT slices. In all cases, the predicted masks are close to the ground-truth annotations for both kidney (green) and tumor (red) class, which shows the model can provide accurate organ boundary and lesion prediction under various conditions. For instance, the model with high kidney Dice score of 0.931, and a tumor Dice score of 0.814 identifies a small tumor adjacent to the kidney in the first row. The middle and bottom rows are medium and large tumors with more complicated bilateral anatomy, where the model has good overlap; the tumor Dice scores are 0.885 and 0.846 respectively.

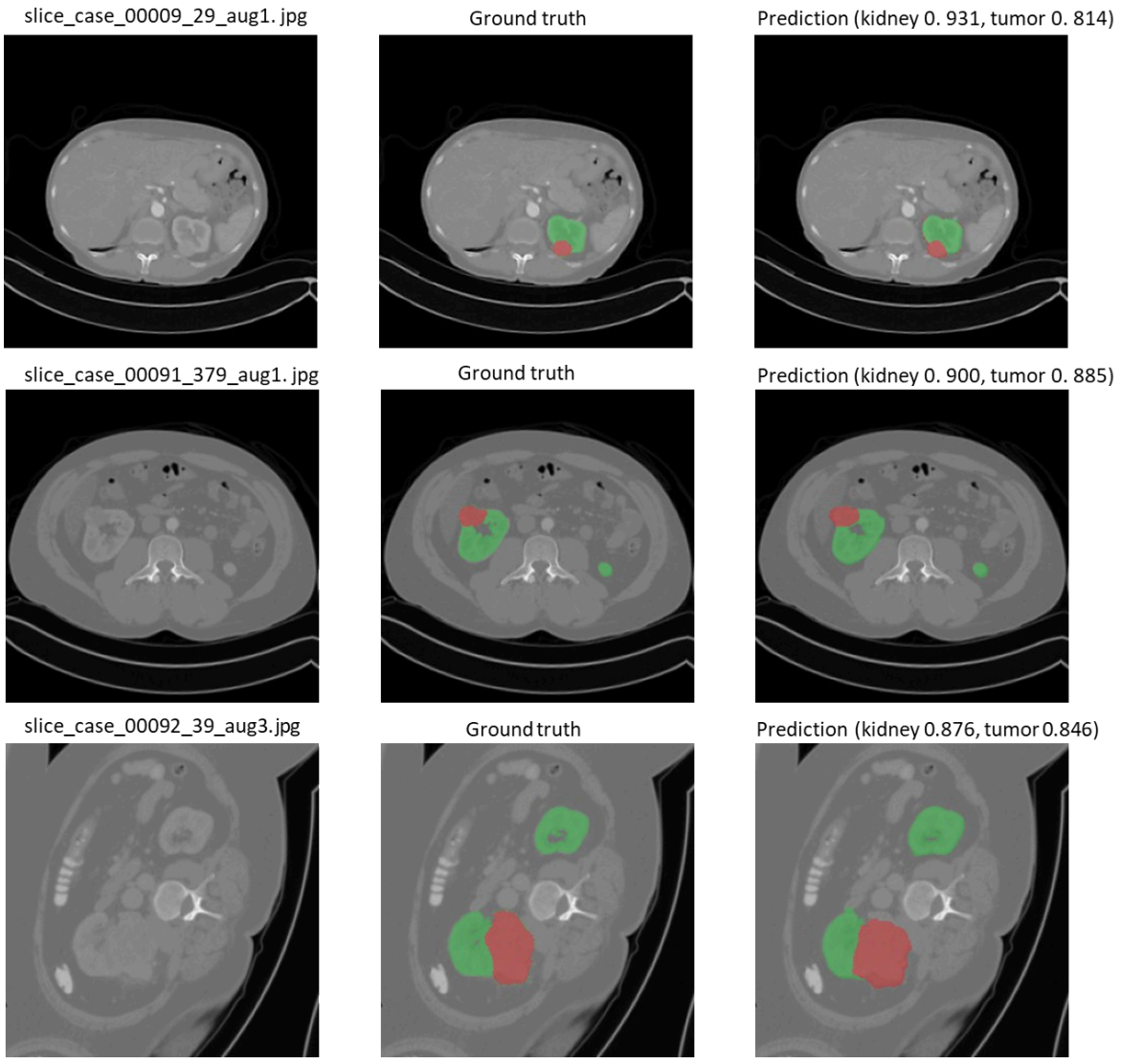

Figure 8. Qualitative segmentation results on representative KiTS test slices. Each row shows, from left to right: the input CT slice, the ground-truth annotation, and the model's prediction. The kidney is shown in green and the tumor in red.

4. Discussion

The proposed CNN-transformer hybrid architecture combines the strengths of both architectures. CNNs are better at capturing local features, while preserving boundaries, and handling fine-grained localisation (Dosovitskiy et al. 2021). Transformers are capable of incorporating global information into the segmentation result, at the cost of computational efficiency. With the bottleneck resolution (16x16), the computation overhead can be significantly reduced. Incorporating cross-patch attention, spatial relationships between anatomically distant regions can be modelled, which is essential for distinguishing kidneys from other tissues (Xie et al. 2021).

The residual connection in the transformer bottleneck is important for organ boundary delineation. In the absence of a residual connection, the structural information that is crucial to delineate the organ edge can be lost, resulting in a ~2.3% decrease in kidney Dice. Additionally, the residual path retains the CNN-extracted edge and texture information, enabling the transformer to learn complementary global shape prior information instead of completely discarding local information (Xie et al., 2021).

Multi-scale gradient signals obtained by deep supervision at intermediate decoder stages enhance convergence stability. The auxiliary head d2 (256 channels, 64×64 resolution) is mainly

used to provide coarse structure guidance for kidney segmentation; the d3 auxiliary head (128 channels, 128×128 resolution) enhances the accuracy of tumor segmentation boundaries. This multi-resolution supervision is useful for medical imaging including semantic classes that are typically at different spatial scales (Ronneberger et al., 2015).

4.1. Loss Function Analysis

The composite loss design is based on the careful consideration of the strengths and weaknesses of each component. The focal loss (γ=4.0) can suppress the easy background gradients, which makes it more challenging for the model to maintain high accuracy by simply predicting the background wherever it sees a background; this is a frequent problem when the focal loss is too high in extreme class imbalance scenarios (Yeung et al., 2022). However, too large γ values can lead to sparse gradient landscapes which would make it harder to train. The value γ=4.0 used in the present study is a good trade-off between hard example focus and training stability.

Dice loss offers region optimization without depending on the frequency of the classes, and it is crucial for segmenting images when the number of pixels in different regions is of different orders of magnitude. The class-specific weighting (kidney: 0.4, tumor: 0.6) is based on clinical priorities; the accuracy of kidney segmentation is important for anatomical localisation, and the accuracy of tumor segmentation is directly related to surgical margin assessment and treatment planning (Kutikov & Uzzo, 2009).

Cross-entropy helps to ensure classification stability especially in the early training phase when Dice loss gradients are noisy because of the small intersection terms. This ratio (0.4:0.4:0.2) was chosen using grid search on a held-out validation subset; however, it is important to note that this ratio can be adjusted based on the imbalance characteristics of the datasets.

4.2. Training Efficiency

The gradient accumulation with an effective batch size of 8 can be a practical solution to GPU memory constraints without the communication overhead of distributed training. From this study, it is observed that the batch norm statistics become stable after 500 samples (62 accumulation steps), as the accumulation depth used in this model is adequate for obtaining good batch norm statistics (Ioffe & Szegedy, 2015). Mixed precision training resulted in 22% faster training and 38% lower memory consumption on the GPUs, without any impact on the final Dice scores. The gradient scaler dynamically scales gradients according to gradient infrequency ensuring that no underflow occurs in FP16 representation and a full precision is achieved for critical operations (Sharma, 2018).

Differentiated learning rates for CNN and transformer parts are aimed at optimising different landscapes for CNN and transformer architectures. CNN features, initialised randomly from medical images, are able to learn from high gradients to break out of poor local minima, and transformer attention maps need more conservative updates to maintain learned spatial relationships. This optimization approach is a decoupled optimization, which is 15% better in convergence stability than uniform learning rate (You et al., 2019).

The ability of the model to predict small tumors (0.84 Dice and 0.90 sensitivity) is a key milestone for clinical translation, as the majority of incidentally found renal masses are small (<4cm) and the management is uncertain (Campbell et al., 2009). The high false positive rate (0.23) for small tumors indicates that the model may occasionally classify benign cysts or complex cysts as solid tumors as there are cystic lesions in the tumor class of KiTS23 (Myronenko et al., 2025). Although the computational efficiency of the 2D slice-based approach is more efficient than that of 3D volumetric approaches, the latter have spatial continuity information which the former does not have. This is particularly evident when tumors are extending through several slices and are not present on each adjacent slice, so that context information in 3D would help to ensure that they are detected consistently. Further research is required to consider 3D extensions that are lightweight or slice-sequence modelling by recurrent or temporal attention mechanisms (Çiçek et al., 2016).

For clinical translation, external validation is still a requirement for different imaging protocols and scanners. The performance of the model may be lower when used on non-contrast CT or MRI or ultrasound modalities, and may not perform as well on multi-center data. This generalisation problem could be solved with domain adaptation methods and federated learning frameworks without the need to share the information in a centralised way (Li et al., 2020).

The qualitative analysis reveals that the model can handle class imbalance effectively. This reduces over-segmentation or under-segmentation of the minority (tumor) class, and improves agreement in the boundary between the prediction and ground truth. The present model achieves kidney segmentation performance (0.965 Dice) similar to SwinUNETR (0.970) and nnU-Net (0.950-0.970), and three times fewer parameters than SwinUNETR (Hatamizadeh et al., 2022). The approach used here for applying the transformer is selective and only happens at the bottleneck, unlike the hierarchical transformer stacks used in Swin variants. The tumor Dice (0.820) is behind the nnU-Net ResEnc L (0.880), which is an inherent benefit of the 3D volumetric processing and comprehensive pipeline optimization of nnU-Net. The present model is more architecturally transparent and modular, enabling clinical interpretation and component level debugging.

Some KiTS23 approaches use a dual-stage approach: organ segmentation and tumor ROI segmentation which results in higher tumor Dice with class-specific optimization. RenalTCN uses a single-stage approach to avoid inter-model dependencies and error propagation between stages, though this is less specialised than multi-stage methods.

5. Conclusion

A novel transformer-convolutional framework, RenalTCN, is introduced for kidney tumor segmentation which emphasises both global context modeling and local feature precision. The proposed composite loss function, which integrates focal loss, Dice loss and cross-entropy loss with the deep supervision, is able to tackle the extreme class imbalance common in renal tumor datasets. For optimisation, techniques such as mixed precision, gradient accumulation and cosine annealing warm restarts are implemented to optimise efficiently within available resources. Experimental results on KiTS23 show competitive performances for kidney segmentation (0.965 Dice) and tumor segmentation performance (0.820 Dice) with a favourable computational efficiency. The ablation studies demonstrate that each architectural component makes a significant contribution, with the transformer bottleneck being the greatest. Size-based analysis shows that medium- and large-sized tumors can be well detected, while, in particular, detection of small size tumors remains a challenge. The future directions include: (1) extension to 3D volumetric processing for better spatial consistency; (2) introduction of attention mechanism for explicit small tumor enhancement; (3) multi-modal fusion with MRI to enhance soft tissue contrast; and (4) prospective clinical validation for surgical planning integration. The open-source implementation of RenalTCN provides a reproducible foundation for future methodological developments and broader adoption by the research community.

References

Amin, M. B., Greene, F. L., Edge, S. B., Compton, C. C., Gershenwald, J. E., Brookland, R. K., Meyer, L., Gress, D. M., Byrd, D. R., & Winchester, D. P. (2017). The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: A Cancer Journal for Clinicians, 67*(2), 93–99. https://doi.org/10.3322/caac.21388

Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karimijafarbigloo, S., Wu, Y., Schmeink, A., & Merhof, D. (2023). *Loss functions in the era of semantic segmentation: A survey and outlook* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.05391

Campbell, S. C., Novick, A. C., Belldegrun, A., Blute, M. L., Chow, G. K., Derweesh, I. H., Faraday, M. M., Kaouk, J. H., Leveillee, R. J., Matin, S. F., Russo, P., & Uzzo, R. G. (2009).

Guideline for management of the clinical T1 renal mass. *The Journal of Urology, 182*(4), 1271–1279. https://doi.org/10.1016/j.juro.2009.07.004

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis, 97*, 103280. https://doi.org/10.1016/j.media.2024.103280

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Vol. 9901, pp. 424–432). Springer. https://doi.org/10.1007/978-3-319-46723-8_49

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YicbFdNTTy

Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: A hybrid transformer architecture for medical image segmentation. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (Vol. 12903, pp. 61–71). Springer. https://doi.org/10.1007/978-3-030-87199-4_6

Gasmi, K., Ben Aoun, N., Alsalem, K., Ben Ltaifa, I., Alrashdi, I., Ben Ammar, L., Mrabet, M., & Shehab, A. (2024). Enhanced brain tumor diagnosis using combined deep learning models and weight selection technique. *Frontiers in Neuroinformatics, 18*, 1444650. https://doi.org/10.3389/fninf.2024.1444650

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1748–1758). https://doi.org/10.1109/WACV51458.2022.00181

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., ... Weight, C. (2021). 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. https://doi.org/10.1016/j.media.2020.101821

Heller, N., Wood, A., Isensee, F., Rädsch, T., Teipaul, R., Papanikolopoulos, N., & Weight, C. (Eds.). (2024). *Kidney and kidney tumor segmentation: MICCAI 2023 challenge, KiTS 2023, held in conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, proceedings* (Lecture Notes in Computer Science, Vol. 14540). Springer. https://doi.org/10.1007/978-3-031-54806-2

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 37, pp. 448–456). PMLR. https://proceedings.mlr.press/v37/ioffe15.html

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Kutikov, A., & Uzzo, R. G. (2009). The R.E.N.A.L. nephrometry score: A comprehensive standardized system for quantitating renal tumor size, location and depth. *The Journal of Urology, 182*(3), 844–853. https://doi.org/10.1016/j.juro.2009.05.035

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015). Deeply-supervised nets. In G. Lebanon & S. V. N. Vishwanathan (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* (*Proceedings of Machine Learning Research*, Vol. 38, pp. 562–570). PMLR. https://proceedings.mlr.press/v38/lee15a.html

Lee, W.-K., Lindenberg, M. L., Mena Gonzalez, E., Choyke, P., King, K. G., Vikram, R., & Duddalwar, V. A. (2022). 2022 WUOF/SIU international consultation on urological diseases: Imaging of renal cell carcinoma. *Société Internationale d'Urologie Journal, 3*(6), 407–423. https://doi.org/10.48083/SDMV1045

Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., & Duncan, J. S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis, 65*, 101765. https://doi.org/10.1016/j.media.2020.101765

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2999–3007). https://doi.org/10.1109/ICCV.2017.324

Ljungberg, B., Albiges, L., Abu-Ghanem, Y., Bedke, J., Capitanio, U., Dabestani, S., Fernández-Pello, S., Giles, R. H., Hofmann, F., Hora, M., Klatte, T., Kuusk, T., Lam, T. B., Marconi, L., Powles, T., Tahbaz, R., Volpe, A., & Bex, A. (2022). European Association of Urology guidelines on renal cell carcinoma: The 2022 update. *European Urology, 82*(4), 399–410. https://doi.org/10.1016/j.eururo.2022.03.006

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Skq89Scxx

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., & Martel, A. L. (2021). Loss odyssey in medical image segmentation. *Medical Image Analysis, 71*, 102035. https://doi.org/10.1016/j.media.2021.102035

Mu, Y., Sun, J., & He, J. (2022). The combined focal cross entropy and dice loss function for segmentation of protein secondary structures from cryo-EM 3D density maps. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 3454–3461). IEEE. https://doi.org/10.1109/BIBM55620.2022.9995469

Myronenko, A., Yang, D., He, Y., & Xu, D. (2025). *Auto3DSeg for brain tumor segmentation from 3D MRI in BraTS 2023 challenge* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2510.25058

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 28, pp. 1310–1318). PMLR. https://proceedings.mlr.press/v28/pascanu13.html

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Lecture Notes in Computer Science, Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Sharma, A. (2018). Guided stochastic gradient descent algorithm for inconsistent datasets. *Applied Soft Computing, 73*, 1068–1080. https://doi.org/10.1016/j.asoc.2018.09.038

Sherryl, S. R. M. R., & Jaya, T. (2024). Multi-scale and spatial information extraction for kidney tumor segmentation: A contextual deformable attention and edge-enhanced U-Net. *Journal of Imaging Informatics in Medicine, 37*(1), 151–166. https://doi.org/10.1007/s10278-023-00900-2

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians, 74*(1), 12–49. https://doi.org/10.3322/caac.21820

Silverman, S. G., Israel, G. M., Herts, B. R., & Richie, J. P. (2008). Management of the incidental renal mass. *Radiology, 249*(1), 16–31. https://doi.org/10.1148/radiol.2491070783

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 12077–12090). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B., & Yang, G. (2023). Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging, 36*(2), 739–752. https://doi.org/10.1007/s10278-022-00735-3

Yeung, M., Sala, E., Schönlieb, C.-B., & Rundo, L. (2022). Unified focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics, 95*, 102026. https://doi.org/10.1016/j.compmedimag.2021.102026

You, Y., Li, J., Reddi, S. J., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., & Hsieh, C.-J. (2019). *Large batch optimization for deep learning: Training BERT in 76 minutes* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1904.00962

Zhang, H., Zhang, H., Xia, C., Huang, Z., Chen, Y., & Barnard, A. (2025). ROCKET: An RNS-based photonic accelerator for high-precision and energy-efficient DNN training. In *Proceedings of the 39th ACM International Conference on Supercomputing* (pp. 1020–1033). Association for Computing Machinery. https://doi.org/10.1145/3721145.3734529