Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Phase-Aware Fusion Framework via Tri-Phase Enhancement Difference Learning for Benign--Malignant Renal Tumor Classification

ruirong wu
Wuyi University
Changyi Ma
Jiangmen Central Hospital
Jin Cui
Jiangmen Central Hospital
Zhixin Luo
Guangdong Medical College
Yunxuan Xie
Guangdong Medical College
Ziqing Wu
Guangdong Medical College
Zhiyong Hong
hongmr@163.com
Wuyi University
Xuechen Li
Wuyi University
Enming Cui
Jiangmen Central Hospital

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11721509/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Phase-Aware Fusion Framework via Tri-Phase Enhancement Difference Learning for Benign–Malignant Renal Tumor Classification

Ruirong Wu[1†], Changyi Ma[2†], Jin Cui[2], Zhixin Luo[3], Yunxuan Xie[3], Ziqing Wu[3], Zhiyong Hong[1*], Xuechen Li[1*], Enming Cui[2*]

[1]School of Electronic and Information Engineering, Wuyi University, Jiangmen, 529020, China.
[2]Department of Radiology, Jiangmen Central Hospital, Jiangmen, 529030, China.
[3]Guangdong Medical University, Zhanjiang, 524000, China.

*Corresponding author(s). E-mail(s): hongmr@163.com; lixuechen1987622@126.com; cem2008@163.com;
†These authors contributed equally to this work.

Abstract

Preoperative differentiation of benign from malignant renal tumors directly informs treatment decisions. However, fat-poor angiomyolipomas, oncocytomas and some renal cell carcinomas (RCCs) have substantially overlapping CT appearances, making them difficult to distinguish by visual assessment. Here, we present a tri-phase enhancement-difference fusion framework guided by phase-level auxiliary supervision to explicitly model enhancement changes across the precontrast phase (PCP), corticomedullary phase (CMP) and nephrographic phase (NP). The proposed UniRepLKNet-based configuration achieved an AUC of 0.981 ± 0.011 and an F1 score of 0.933 ± 0.022 in the representative-method comparison. Five-seed ablation experiments also assessed the individual and combined contributions of TEDM and PAS to model performance. To our knowledge, this is the first end-to-end framework for renal tumor differentiation that explicitly models three-phase enhancement differences and incorporates phase-level auxiliary supervision during training. This study provides a comparative baseline for multiphase renal tumor differentiation and may inform the development of more efficient phase-aware approaches.

Keywords: renal tumors, deep learning, multiphase CT, enhancement-difference fusion

1 Introduction

With the widespread use of abdominal computed tomography (CT), renal tumors, particularly incidentally detected solid enhancing masses, are being identified with increasing frequency. Preoperative differentiation of benign from malignant disease is therefore a key determinant of management. Yet renal cell carcinoma (RCC) and benign tumors such as fat-poor angiomyolipoma and oncocytoma often show overlapping morphology and enhancement on contrast-enhanced CT, making reliable visual discrimination difficult. Diagnosis consequently continues to rely on

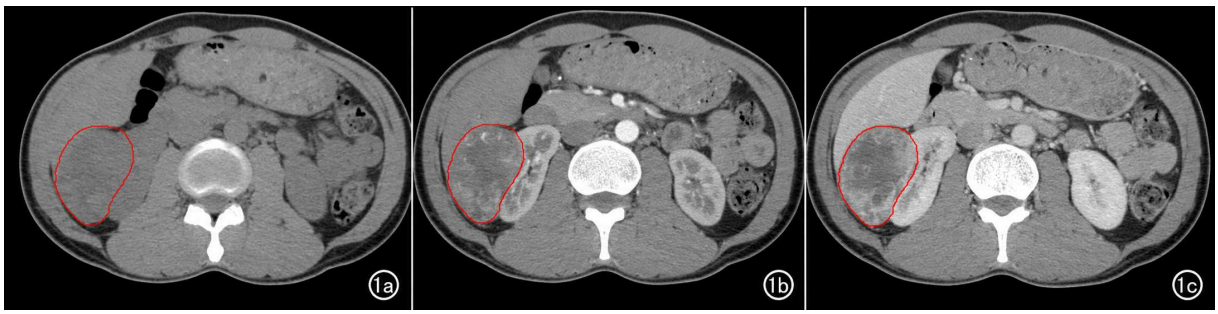

Fig. 1 Contrast-enhanced CT appearances of a malignant renal tumor across three phases. (a) Precontrast phase; (b) corticomedullary phase; (c) nephrographic phase.

percutaneous biopsy or surgical histopathology [1, 2]. This uncertainty may lead to unnecessary resection of benign masses suspected to be malignant and exposes patients to the risks of invasive diagnosis. Reliable, non-invasive preoperative differentiation therefore has clear clinical value.

Multiphase contrast-enhanced CT captures the time-dependent enhancement kinetics of contrast material rather than morphology at a single time point, and is widely used to characterize renal tumors. Previous studies have demonstrated the value of multiphase information through conventional imaging, radiomics and deep learning [3–5].

Young et al. compared enhancement patterns across precontrast, corticomedullary, nephrographic and excretory phases in clear-cell, papillary and chromophobe RCC and oncocytoma. They found phase-specific differences in peak enhancement and showed that multiphase enhancement thresholds could distinguish some histological subtypes, supporting the diagnostic value of cross-phase enhancement trajectories [7]. Dai et al. developed a deep-learning algorithm based on contrast-enhanced multiphase CT to estimate the likelihood that a small renal mass was benign. Validation in internal, external and prospective cohorts suggested that multiphase CT can provide automated support for benignity assessment [4]. Xiong et al. further used preoperative multiphase CT to predict renal tumor malignancy, aggressive pathological features and survival, suggesting that multiphase CT representations may connect benign–malignant classification with pathology and prognosis [8]. More recently, Uhm et al. introduced a lesion-aware cross-phase attention network that treated multiphase CT as a temporal sequence and learned dependencies among enhancement phases for five-class renal tumor subtype classification. They also noted that many earlier deep-learning methods relied on early fusion, simply concatenating images from different phases along the channel dimension without explicitly modeling interphase enhancement relationships [9]. In our earlier work on the same cohort, the three-phase model yielded an area under the receiver operating characteristic curve (AUC) of 0.84, exceeding the AUCs of individual phases (0.77–0.79) [6].

An important gap therefore remains. Existing multiphase methods have mainly established the value of joint phase input or learned interphase dependencies in a data-driven manner through attention. However, they do not directly and efficiently model radiologically meaningful differences among PCP, CMP and NP, including early enhancement (CMP relative to PCP), cumulative enhancement relative to the precontrast baseline (NP relative to PCP) and delayed enhancement change (NP relative to CMP). For example, hypervascular clear-cell RCC often shows marked enhancement during the corticomedullary phase followed by subsequent changes, whereas some fat-poor angiomyolipomas and oncocytomas may exhibit different enhancement trajectories. Meanwhile, models optimized solely with the final classification loss may rely disproportionately on the most discriminative single phase and underuse complementary multiphase information.

To address this gap, we propose a tri-phase enhancement-difference fusion framework guided by phase-level auxiliary supervision. The main contributions are as follows:

1. We introduce a tri-phase enhancement-difference module (TEDM) that explicitly models early, cumulative and delayed enhancement changes at the feature level rather than relying on implicit learning from pixel-level inputs. It then combines these changes through spatially adaptive fusion with residual compensation.
2. We introduce a phase-level auxiliary supervision (PAS) strategy that applies a shared auxiliary head to each phase before feature fusion. This encourages each phase to retain a discriminative representation and discourages shortcut reliance on the single most discriminative phase. The auxiliary heads are removed at inference, adding no inference-time computational cost.

2 Related work

2.1 Radiomics and machine learning for benign–malignant differentiation of renal tumors

Radiomics combined with conventional machine learning remains an established approach to differentiating benign from malignant renal tumors. It typically extracts high-dimensional hand-crafted features, including shape, texture and enhancement descriptors, from delineated regions of interest (ROIs), selects features statistically and then passes them to a classifier. A recent multicenter study trained random-forest models using individual phases and all three phases of contrast-enhanced CT and found better discrimination with the three-phase input [3]. For the particularly challenging distinction between fat-poor angiomyolipoma and clear-cell RCC, a systematic review and a nomogram study also suggest that CT radiomics can improve on visual assessment [1, 2]. In our earlier work, a two-stage pipeline combining ResNet-50 feature extraction, Mann–Whitney feature selection and an extreme learning machine classifier achieved an AUC of 0.84 with the three-phase data used here [6].

Although the discriminative value of this two-stage paradigm is well established, it has inherent limitations. Hand-crafted or semi-hand-crafted features are decoupled from the classifier, their representation is constrained by predefined operators, and the pipeline is not readily suited to end-to-end learning of dynamic interphase relationships. These limitations motivate further investigation of end-to-end approaches.

2.2 Deep learning for renal tumor classification

Deep learning removes the need for predefined radiomic features and can learn high-dimensional discriminative representations directly from CT images. It has therefore been increasingly applied to benign–malignant differentiation and histological subtype classification of renal tumors. Klontzas et al. transferred ImageNet-pretrained convolutional neural networks to benign–malignant classification of renal tumors on contrast-enhanced CT. They fine-tuned VGG-16, InceptionV3 and Inception-ResNetV2 on an international multicenter dataset spanning multiple institutions and scanner manufacturers, and further evaluated performance using ensemble consensus. Inception-ResNetV2 achieved relatively strong discrimination, suggesting the potential of deep learning for this task; nevertheless, the study emphasized that distinguishing benign from malignant renal tumors on CT remains challenging [5]. For multiphase CT, Dai et al. developed a deep-learning algorithm to estimate the likelihood of benignity in small renal masses. Using precontrast, corticomedullary and nephrographic images, they assessed the masses for malignancy and validated model performance in internal, external and prospective datasets, demonstrating the potential of multiphase CT deep learning for preoperative assessment [4].

Overall, current deep-learning approaches have largely advanced along two directions: stronger backbones and multiphase input. Some use image- or feature-level concatenation for early fusion, whereas others use attention mechanisms to learn interphase dependencies in a data-driven manner. The former do not explicitly represent enhancement trajectories, whereas the latter are typically more complex and may be more sensitive to training-set size. Here, our methodological contribution centers on feature-level modeling of radiologically meaningful enhancement differences among PCP, CMP and NP, with phase-level auxiliary supervision constraining phase representations before fusion. This provides a lightweight

and potentially interpretable approach to using multiphase dynamics.

2.3 Visual backbones and difference-based fusion

At the backbone level, state-space models have emerged as a third architectural family alongside convolutional networks and Transformers. Mamba uses selective state-space models for linear-complexity long-sequence modeling [10]; Vim introduces bidirectional state-space modeling for visual representation learning [11]; VMamba builds a linear-complexity vision backbone using two-dimensional selective scanning [12]; and MambaVision combines self-attention with state-space modeling to balance accuracy and throughput [13]. In medical image classification, MedMamba combines convolutional and state-space components [14], whereas MedViT uses a convolution–Transformer hybrid to improve robustness and generalization [15]. We use UniRepLKNet [19], a large-kernel convolutional network, as the shared visual encoder.

Multiphase renal-CT studies have explored several ways to combine phase information. Han et al. combined weighted three-phase regions of interest and concatenated them as inputs to a CNN for renal tumor subtype classification [20]. Uhm et al. aligned multiphase tumor images and used a three-channel input for end-to-end subtype classification [21]. Dai et al. used a multiview CNN to fuse phase-specific features for benign-probability assessment of small renal masses [4]. Later work introduced lesion-aware cross-phase attention to model phase dependencies [9], while Luo et al. used phase attention to aggregate CT features before fusing them with clinical information [22]. Together, these studies illustrate early image fusion, feature aggregation, and attention-based cross-phase fusion. However, the cited deep-learning models do not encode CMP–PCP, NP–PCP, and NP–CMP as separate feature-level differences. Our TEDM explicitly models these differences, while PAS provides phase-level auxiliary supervision before fusion.

3 Methods

3.1 Overall framework

The framework comprises three components: shared encoding of the three phases, explicit fusion of enhancement differences, and phase-level auxiliary supervision during training. The PCP, CMP and NP images are independently mapped to phase-specific features by a shared-weight encoder. TEDM then explicitly computes early, cumulative and delayed enhancement changes and combines them through spatially adaptive selection and residual fusion to obtain discriminative features. A classification head outputs the probability of malignancy. During training, a PAS branch is added to encourage discriminative representations for each phase. This branch is removed in its entirety at inference and therefore adds no inference-time computational cost. The overall architecture is shown in Fig. 2.

3.2 Shared-weight encoding of the three phases

For a given case, let the input from phase ϕ be:

$$\mathcal{I}_\phi = \{i_\phi^1, i_\phi^2, \ldots, i_\phi^K\}, \qquad \phi \in \{p, c, n\}. \tag{1}$$

Here, p, c and n correspond to PCP, CMP and NP, respectively, and K denotes the number of slices in each phase. Each slice i_ϕ^j is passed through a shared encoder E_θ to extract spatial features:

$$f_\phi^j = E_\theta(i_\phi^j), \qquad f_\phi^j \in \mathbb{R}^{C\times h\times w}. \tag{2}$$

The features from the K slices within each phase are then averaged along the anatomical-slice dimension to obtain a case-level phase representation:

$$F_\phi = \frac{1}{K}\sum_{j=1}^{K} f_\phi^j, \qquad F_\phi \in \mathbb{R}^{C\times h\times w}. \tag{3}$$

All three phases share a single parameter set θ rather than using three independent branches. This constraint maps the phases into a common representation space, so subsequent differences between phase features capture enhancement changes rather than parameter offsets across

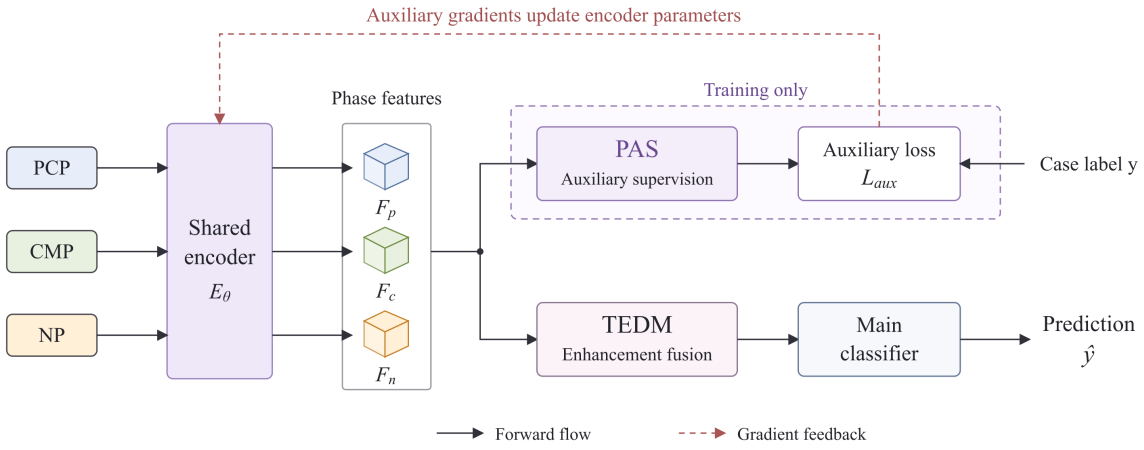

Fig. 2 Overview of the proposed framework.

branches. Parameter sharing also reduces the number of parameters and acts as regularization in this small-sample setting.

3.3 Tri-phase enhancement-difference module (TEDM)

As shown in Fig. 3, the module first explicitly decomposes enhancement dynamics into three relationships:

$$\begin{aligned} D_1 &= F_c - F_p, \\ D_2 &= F_n - F_p, \\ D_3 &= F_n - F_c. \end{aligned} \tag{4}$$

Here, D_1 represents early enhancement, D_2 cumulative enhancement, and D_3 delayed change. Each difference feature is transformed using a shared difference encoder ϕ:

$$\begin{aligned} U_i &= W_1(\mathrm{LN}(D_i)), \\ V_i &= \mathrm{DWConv}(U_i), \\ Q_i &= \mathrm{GELU}(V_i), \\ E_i = \phi(D_i) &= W_2(\mathrm{GRN}(Q_i)), \qquad i \in \{1,2,3\}. \end{aligned} \tag{5}$$

Here, U_i, V_i, and Q_i denote the normalized projection, depthwise-convolved feature, and activated feature, respectively. LN denotes layer normalization, and DWConv denotes depthwise convolution, which provides local spatial interactions. GRN (global response normalization) promotes competition among channel responses, making salient channels in the enhancement-difference features easier for the fusion module to select. W_1 and W_2 are pointwise linear projections: W_1 maps C channels to a hidden dimension C_1, and W_2 projects the features back to C channels so that E_i has the same dimensionality as the original phase features. All three difference types share ϕ, not to erase their distinctions, but to provide a common, stable base representation and reduce the risk of overfitting associated with three independent encoders in a small dataset. Phase-specific differences are instead modeled during fusion by the subsequent spatial-selection mechanism.

After computing the three difference representations, the module performs spatial selection and residual fusion. Enhancement patterns are not equally informative throughout a tumor: some regions may be characterized by early enhancement, whereas others may require delayed changes in the nephrographic phase. We therefore introduce a lightweight spatial-selection mechanism. It concatenates the original three-phase features with the three difference representations along the channel dimension, then uses a pointwise convolution ψ to predict spatial weights for the three differences, followed by Softmax normalization:

$$\begin{aligned} Z &= \mathrm{concat}[F_p, F_c, F_n, E_1, E_2, E_3], \\ [A_1, A_2, A_3] &= \mathrm{Softmax}(\psi(Z)), \end{aligned} \tag{6}$$

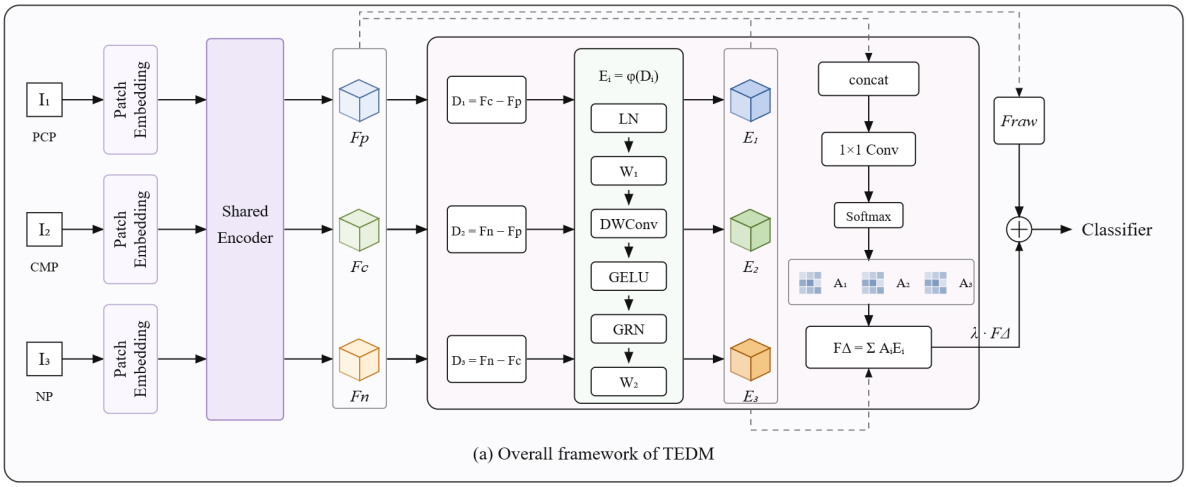

Fig. 3 Overall architecture of TEDM.

$$F_\Delta = A_1 \odot E_1 + A_2 \odot E_2 + A_3 \odot E_3. \quad (7)$$

Compared with computationally heavier cross-attention or cross-phase Transformers, this selection mechanism uses substantially fewer parameters and is better suited to regularization in small datasets. To retain absolute morphology, boundaries and background information that may be lost if only differences are used, we preserve a residual pathway from the original three-phase features and introduce a learnable scalar λ to balance the original image information and enhancement dynamics:

$$\begin{aligned} F_{\text{raw}} &= W_{\text{raw}}\,(\text{LN}([F_p, F_c, F_n]))\,, \\ F_{\text{out}} &= F_{\text{raw}} + \lambda\, F_\Delta. \end{aligned} \quad (8)$$

The fused features are passed through global average pooling and a compact classification head. Let u denote the normalized pooled representation and v the activated hidden representation:

$$\begin{aligned} u &= \text{LN}(\text{GAP}(F_{\text{out}}))\,, \\ v &= \text{GELU}(W_1^c(u))\,, \\ \hat{y} &= W_2^c(\text{Dropout}(v))\,. \end{aligned} \quad (9)$$

3.4 Phase-level auxiliary supervision (PAS)

As shown in Fig. 4, we use a single shared auxiliary classification head to generate phase-level predictions from the features for the three phases. The head comprises global average pooling, layer normalization and a linear classifier [16]. Let u_ϕ denote the normalized pooled feature for phase ϕ:

$$\begin{aligned} u_\phi &= \text{LN}(\text{GAP}(F_\phi))\,, \\ g(F_\phi) &= W_{\text{aux}}(u_\phi), \\ \hat{y}_\phi^{\text{aux}} &= g(F_\phi), \qquad \phi \in \{p, c, n\}. \end{aligned} \quad (10)$$

The phase-level auxiliary loss is defined as:

$$\mathcal{L}_{\text{aux}} = \frac{1}{3} \sum_{\phi \in \{p,c,n\}} \ell_{\text{CE}}(\hat{y}_\phi^{\text{aux}}, y). \quad (11)$$

Here, y is the ground-truth label for a case, and ℓ_{CE} denotes the cross-entropy loss. The auxiliary head is optimized only during training. At inference, the model uses only the main classification head after TEDM fusion to produce the final benign–malignant prediction, so the auxiliary branch adds no computational cost at inference.

3.5 Total loss function

The model is optimized jointly using the main classification loss and the phase-level auxiliary loss. Because the benign and malignant classes are imbalanced in this study, standard cross-entropy may be dominated by the majority class. We therefore use focal loss in the main classification branch to increase emphasis on minority-class and difficult examples [17]:

$$\mathcal{L}_{\text{main}} = -\alpha(1 - p_t)^\gamma \log(p_t). \quad (12)$$

Here, p_t denotes the model-predicted probability of the true class, α is the class-balancing factor,

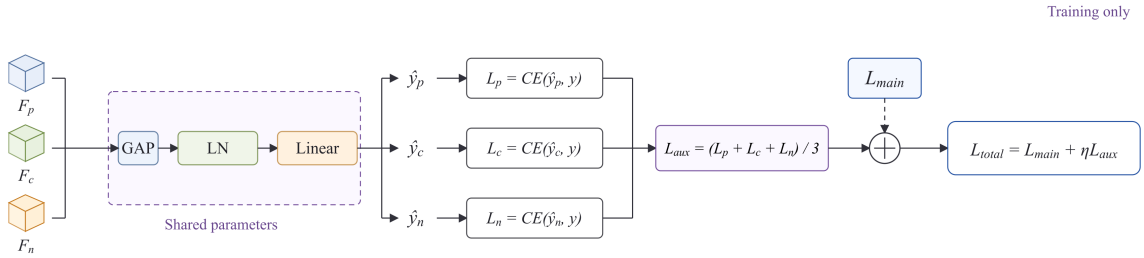

Fig. 4 Overview of PAS.

and γ is the focusing parameter. When an example is correctly classified and p_t is high, $(1-p_t)^\gamma$ reduces its loss weight; difficult examples retain a larger gradient contribution.

The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \eta \mathcal{L}_{\text{aux}}. \tag{13}$$

Here, η weights the auxiliary loss and balances the final fused-classification and phase-level auxiliary supervision.

4 Experiments

4.1 Dataset

The dataset comprised an in-house cohort and external data from two independent institutions and The Cancer Imaging Archive (TCIA). The in-house cohort comprised 496 lesions (418 malignant and 78 benign) and was used for training and internal validation. An independent external test set comprised 309 lesions (262 malignant and 47 benign) from external institutions and TCIA. The institutional review board of our hospital approved the study and waived the requirement for informed consent.

4.2 Implementation details

Original Digital Imaging and Communications in Medicine (DICOM) images were converted to NIfTI format using MRIcroGL. An experienced abdominal radiologist delineated tumor regions of interest slice by slice using ITK-SNAP. The three phases were registered in advance so that corresponding voxels in PCP, CMP and NP represented the same anatomical locations. For each lesion, we selected five contiguous anatomical slices: the slice with the largest region of interest and the two slices above and below it. Images were displayed using a soft-tissue window (window level, 40 Hounsfield units (HU); window width, 400 HU) to preserve enhancement-related attenuation information. During augmentation, identical geometric transformations (random flips and rotations) were applied to all slices and phases from the same case to preserve spatial correspondence across phases.

We used a batch size of 16 and AdamW [18] for optimization, with a cosine-annealing learning-rate schedule and a dropout rate of 0.6 in the classification head. Structural hyperparameters for TEDM were selected using the internal validation set; the external test set was not used for hyperparameter or threshold selection. Tables 2 and 3 report the mean ± standard deviation across five random seeds; Fig. 5 visualizes mean performance across backbones. All experiments were conducted on a single NVIDIA RTX 5090 GPU.

4.3 Comparative experiments

For historical context, Table 1 reports the single-phase and three-phase results from our earlier study on this cohort [6]. The three-phase input (All) achieved an AUC of 0.84, higher than the AUCs for PCP (0.77), CMP (0.78) and NP (0.79), an absolute difference of 0.05 from the best-performing single phase, NP. These previously reported point estimates are separate from the five-seed comparisons in Tables 2 and 3.

Table 2 compares representative model configurations with the complete proposed method, in which "ours" denotes the UniRepLKNet backbone combined with TEDM and PAS. The proposed configuration had the highest mean AUC (0.981 ± 0.011), accuracy (0.898 ± 0.036), sensitivity (0.986 ± 0.024) and F1 score (0.933 ± 0.022) among the methods listed.

Table 1 Single-phase versus three-phase input

Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
PCP	0.77	0.66	0.65	**0.84**	0.93	0.27
CMP	0.78	0.73	0.77	0.53	0.9	0.29
NP	0.79	**0.78**	**0.79**	0.7	0.93	**0.38**
All	**0.84**	0.76	0.75	0.80	**0.95**	0.36

Table 2 Comparison with representative methods

Model	AUC	Acc	Sen	F1
MedViTV2	0.899 ± 0.010	0.792 ± 0.040	0.717 ± 0.073	0.828 ± 0.039
MedMamba	0.856 ± 0.024	0.763 ± 0.042	0.711 ± 0.092	0.806 ± 0.049
MambaVision	0.937 ± 0.016	0.865 ± 0.026	0.867 ± 0.059	0.900 ± 0.023
FusionMamba	0.906 ± 0.020	0.843 ± 0.033	0.867 ± 0.045	0.886 ± 0.026
ours	**0.981 ± 0.011**	**0.898 ± 0.036**	**0.986 ± 0.024**	**0.933 ± 0.022**

Table 3 Component ablation for the UniRepLKNet-based model

Configuration	AUC	Accuracy	Sensitivity	F1
Baseline	0.945 ± 0.058	0.869 ± 0.040	0.975 ± 0.037	0.913 ± 0.027
+ TEDM	0.965 ± 0.049	**0.898 ± 0.068**	0.961 ± 0.054	0.931 ± 0.045
+ TEDM + PAS	**0.981 ± 0.011**	**0.898 ± 0.036**	**0.986 ± 0.024**	**0.933 ± 0.022**

4.4 Ablation experiments

4.4.1 Component ablation

The component ablation in Table 3 was conducted using UniRepLKNet as the backbone. Compared with the baseline, adding TEDM increased the mean AUC, accuracy, and F1 score from 0.945 ± 0.058, 0.869 ± 0.040, and 0.913 ± 0.027 to 0.965 ± 0.049, 0.898 ± 0.068, and 0.931 ± 0.045, respectively. The increase in AUC indicates improved discrimination between benign and malignant tumors, while the concurrent gains in accuracy and F1 score show that this improvement was reflected in the overall classification performance. Sensitivity, however, decreased slightly from 0.975 ± 0.037 to 0.961 ± 0.054 with TEDM alone. Adding PAS to TEDM further increased the mean AUC from 0.965 ± 0.049 to 0.981 ± 0.011, sensitivity from 0.961 ± 0.054 to 0.986 ± 0.024, and F1 score from 0.931 ± 0.045 to 0.933 ± 0.022, while accuracy remained at 0.898. The increase in sensitivity above the baseline level suggests that PAS complements TEDM by strengthening malignant-case detection. Overall, the combined TEDM+PAS configuration achieved higher mean values across all four metrics than the baseline, supporting the benefit of combining the two components.

4.4.2 Backbone ablation

Figure 5 summarizes the mean AUC, accuracy, sensitivity, and F1 score for the baseline, TEDM, and TEDM+PAS configurations across five backbones. Relative to each baseline, the TEDM+PAS profile extends outward along all four metrics for every evaluated architecture, showing that the gains observed with UniRepLKNet were also present with FusionMamba, MedMamba, MambaVision, and MedViTV2. The TEDM-only profiles are less uniform: sensitivity decreases slightly for MedMamba (0.711 to 0.708) and UniRepLKNet (0.975 to 0.961), whereas adding PAS raises sensitivity above baseline in both cases. These results indicate that the observed gains are not confined to a single backbone and suggest that the method is compatible with the architectures evaluated here.

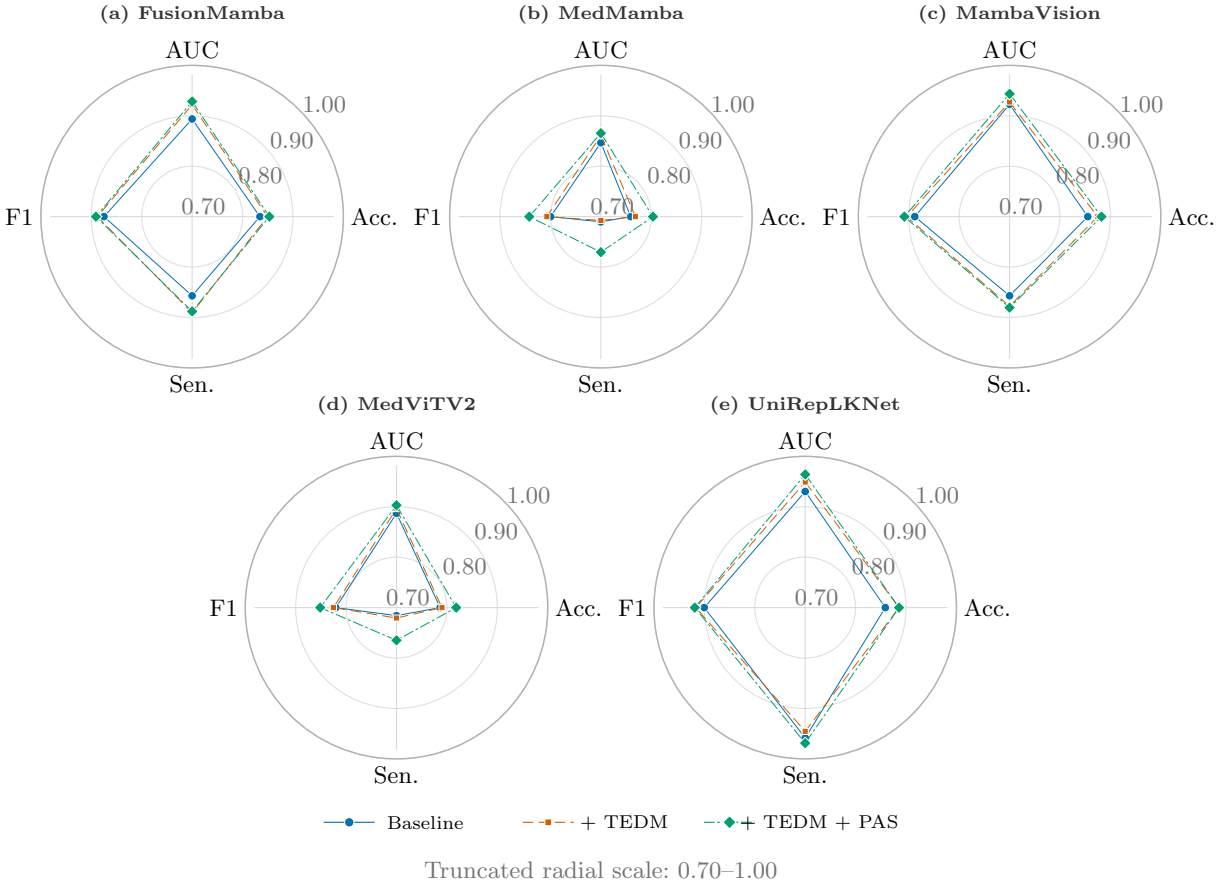

Fig. 5 Backbone ablation across five architectures.

4.5 Visualization analysis

To compare spatial responses before and after explicit modeling of enhancement differences, Fig. 6 shows response heat maps for a representative case across PCP, CMP and NP. Compared with the baseline, the model incorporating TEDM showed responses more concentrated within the lesion in the PCP and NP images. This pattern suggests that explicit difference modeling may help the model focus on regions associated with lesion enhancement changes.

5 Discussion

We developed a tri-phase enhancement-difference fusion framework guided by phase-level auxiliary supervision. The framework introduces enhancement kinetics from three-phase contrast-enhanced CT into a deep-learning model through explicit difference modeling and phase-level auxiliary supervision. Across the five-seed representative-method comparison, the complete UniRepLKNet-based configuration achieved the highest mean AUC, accuracy, sensitivity and F1 score among the listed methods.

Explicit difference modeling provides a radiologically motivated alternative to implicit fusion. Conventional implicit fusion concatenates images or features from all three phases and leaves the network to infer which interphase changes matter. TEDM instead constructs three cross-phase differences corresponding to early enhancement, cumulative enhancement and delayed change. This encodes radiologically meaningful phase relationships as a structural prior and makes the modeled differences explicit. The present experiments do not establish whether this design reduces the

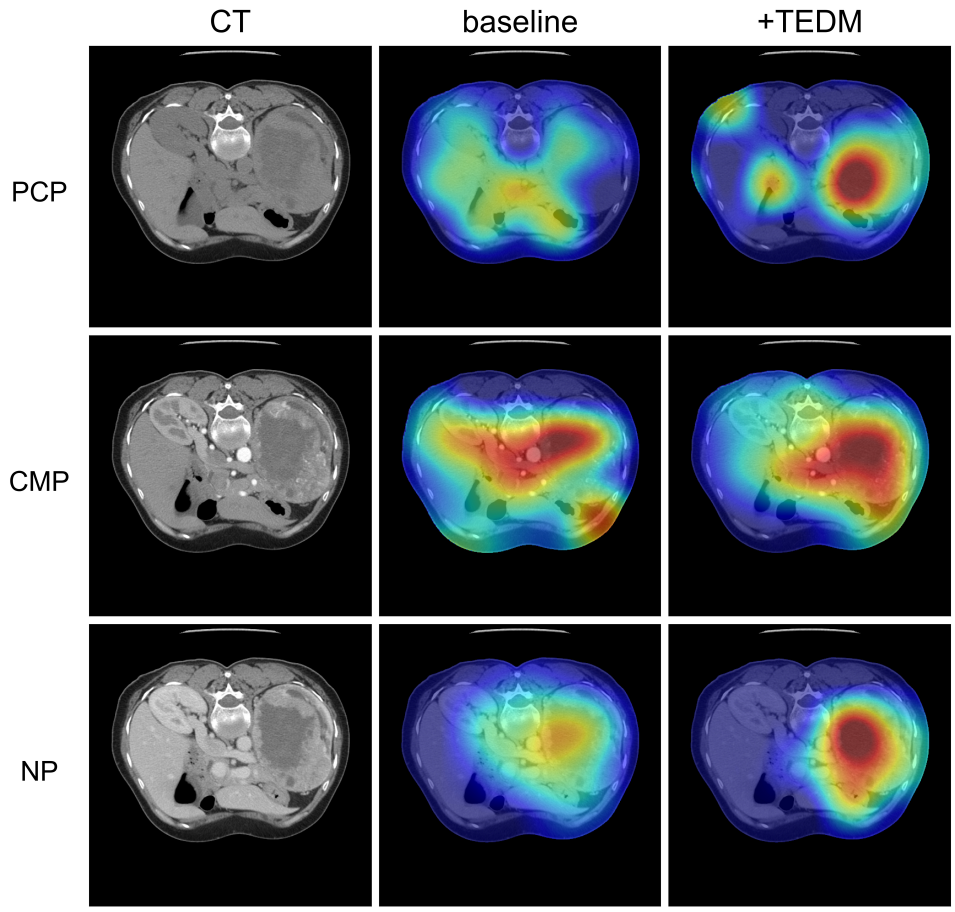

Fig. 6 Representative response heat maps across three-phase CT before and after adding TEDM.

amount of training data required or improves clinical interpretability; these effects warrant direct evaluation.

Several limitations remain.

First, cohort size and class distribution constrain the precision and statistical power of the estimates. Benign lesions were relatively few (78 in the internal cohort and 47 in the external cohort) and markedly outnumbered by malignant lesions. More robust validation in larger, multicenter and prospective cohorts is needed.

Second, the range of tumor types and comparison groups remains limited. The benign cohort included only fat-poor angiomyolipoma and oncocytoma; it does not represent the full spectrum of benign renal tumors, and lesions were not further stratified by subtype. For lesions with weak or atypical enhancement changes, cross-phase difference signals may be weak, and misclassification may persist.

6 Conclusion

This study addresses benign–malignant differentiation of renal tumors on three-phase contrast-enhanced CT. To address implicit fusion of multiphase information and the lack of explicit modeling of interphase enhancement changes, we propose a tri-phase enhancement-difference fusion framework guided by phase-level auxiliary supervision. TEDM explicitly models and spatially adaptively fuses early, cumulative and delayed enhancement changes. PAS uses shared phase-level auxiliary supervision to encourage each phase to remain individually discriminative and to discourage shortcut reliance on a single phase, without adding inference-time computation. Together, the two components link phase-level predictions

with phase-specific feature fusion to form a coherent pipeline. In the representative-method comparison, the complete UniRepLKNet-based configuration achieved mean AUC and F1 scores of 0.981 ± 0.011 and 0.933 ± 0.022 across five random seeds. In the input-phase comparison, the three-phase model had a higher AUC point estimate than any single-phase model.

The study has limitations. First, the pathological subtypes included here were not analyzed separately. Second, the number of benign lesions was small and the sources of external testing data were limited, constraining statistical power. Future studies should evaluate the framework in larger multicenter datasets and extend the analysis to subtype-stratified and prospective settings.

Statements and Declarations

Funding. This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2024A1515220079, 2023A1515220156, and 2025A1515140027), the Jiangmen Basic and Applied Basic Research Project (Grant No. 2520002000123), and the Shenzhen Science and Technology Program (Grant Nos. JCYJ20240813112420027, JCYJ20230807093820041, and JCYJ20220530155811025).

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. This retrospective study was approved by the Ethics Committee of Jiangmen Central Hospital (approval number: Ethics Review [2023]05).

Consent to participate. The requirement for informed consent was waived by the ethics committee owing to the retrospective nature of the study.

Consent for publication. Not applicable.

Data availability. The raw datasets generated and/or analyzed during the current study contain sensitive patient information and are not publicly available due to privacy restrictions mandated by the Ethics Committee of Jiangmen Central Hospital. De-identified summary data may be available from the corresponding authors on reasonable request, subject to approval by the ethics committee.

Code availability. The code used in this study is available from the corresponding authors on reasonable request.

Author contributions. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ruirong Wu, Changyi Ma, Zhiyong Hong, Xuechen Li, Enming Cui, Jin Cui, Zhixin Luo, Yunxuan Xie, and Ziqing Wu. The first draft of the manuscript was written by Ruirong Wu and Changyi Ma, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Dehghani Firouzabadi F, Gopal N, Hasani A, et al. CT radiomics for differentiating fat poor angiomyolipoma from clear cell renal cell carcinoma: systematic review and meta-analysis. PLoS One, 2023; 18(7): e0287299. https://doi.org/10.1371/journal.pone.0287299

[2] Xiao Y, Yang M, Jiang Z, et al. A nomogram model for challenging cases: differentiating fat-poor angiomyolipoma from clear cell renal cell carcinoma in uncertain or misdiagnosed tumors. Abdominal Radiology, 2026; 51(8): 4008-4017. https://doi.org/10.1007/s00261-025-05362-4

[3] Wu Y, Cao F, Lei H, et al. Interpretable multiphasic CT-based radiomic analysis for preoperatively differentiating benign and malignant solid renal tumors: a multicenter study. Abdominal Radiology, 2024; 49: 3096-3106. https://doi.org/10.1007/s00261-024-04351-3

[4] Dai C, Xiong Y, Zhu P, et al. Deep Learning Assessment of Small Renal Masses at Contrast-enhanced Multiphase CT. Radiology, 2024; 311(2): e232178. https://doi.org/10.1148/radiol.232178

[5] Klontzas ME, Kalarakis G, Koltsakis E, et al. Convolutional neural networks for the differentiation between benign and malignant renal tumors with a multicenter international computed tomography dataset. Insights into Imaging, 2024; 15(1): 26. https://doi.org/10.1186/s13244-023-01601-8

[6] CHEN M, ZHOU T, CHEN X, et al. Deep learning model based on CT images for differentiating benign and malignant renal tumors. Radiologic Practice, 2023, 38(4): 468-473. https://doi.org/10.13609/j.cnki.1000-0313.2023.04.016

[7] Young JR, Margolis D, Sauk S, et al. Clear cell renal cell carcinoma: discrimination from other renal cell carcinoma subtypes and oncocytoma at multiphasic multidetector CT. Radiology, 2013; 267(2): 444-453. https://doi.org/10.1148/radiol.13112617

[8] Xiong Y, Yao L, Lin J, et al. Artificial intelligence links CT images to pathologic features and survival outcomes of renal masses. Nature Communications, 2025; 16: 1425. https://doi.org/10.1038/s41467-025-56784-z

[9] Uhm K-H, Jung S-W, Hong S-H, Ko S-J. Lesion-aware cross-phase attention network for renal tumor subtype classification on multi-phase CT scans. Computers in Biology and Medicine, 2024; 178: 108746. https://doi.org/10.1016/j.compbiomed.2024.108746

[10] Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv, 2023. https://arxiv.org/abs/2312.00752

[11] Zhu L, Liao B, Zhang Q, et al. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. Proceedings of the 41st International Conference on Machine Learning, 2024. https://arxiv.org/abs/2401.09417

[12] Liu Y, Tian Y, Zhao Y, et al. VMamba: Visual State Space Model. Advances in Neural Information Processing Systems, 2024. https://arxiv.org/abs/2401.10166

[13] Hatamizadeh A, Kautz J. MambaVision: A Hybrid Mamba-Transformer Vision Backbone. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025. https://arxiv.org/abs/2407.08083

[14] Yue Y, Li Z. MedMamba: Vision Mamba for Medical Image Classification. arXiv, 2024. https://arxiv.org/abs/2403.03849

[15] Manzari ON, Ahmadabadi H, Kashiani H, et al. MedViT: A Robust Vision Transformer for Generalized Medical Image Classification. Computers in Biology and Medicine, 2023; 157: 106791. https://doi.org/10.1016/j.compbiomed.2023.106791

[16] Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-Supervised Nets. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics. PMLR, 2015; 38: 562-570. https://proceedings.mlr.press/v38/lee15a.html

[17] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[18] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. International Conference on Learning Representations, 2019. https://arxiv.org/abs/1711.05101

[19] Ding X, Zhang Y, Ge Y, Zhao S, Song L, Yue X, et al. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio, Video, Point Cloud, Time-Series and Image Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 5513–5524.

[20] Han S, Hwang SI, Lee HJ. The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method. Journal of Digital Imaging, 2019; 32: 638–643. https://doi.org/10.1007/s10278-019-00230-2

[21] Uhm K-H, Jung S-W, Choi MH, et al. Deep learning for end-to-end kidney cancer

diagnosis on multi-phase abdominal computed tomography. npj Precision Oncology, 2021; 5: 54. https://doi.org/10.1038/s41698-021-00195-y

[22] Luo S, Quan Q, Yang R, et al. Multimodal deep learning model for multiclass classification of renal tumors. npj Digital Medicine, 2026; 9: 524. https://doi.org/10.1038/s41746-026-02697-0