UNLOCKING CLINICAL POTENTIAL: BEYOND SINGLE-TO-TRI-PHASE CT WITH DYNAMIC FUSION FOR LIVER TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Liver tumor segmentation is essential for treatment planning and disease monitoring. Most existing methods rely on single-phase computed tomography (CT), they often suffer from low contrast and incomplete lesion depiction. Contrast-Enhanced CT (CECT) offers multiple imaging phases: arterial (ART), portal venous (PV), and delayed (DL), which provide complementary anatomical and functional information. This study begins with a systematic quantitative evaluation of each enhanced phase using standard segmentation models to investigate their individual contributions and validate phase-specific clinical insights. Guided by this analysis, a Multi-phase Attention Deep Fusion Network (`MADF-Net`) is proposed to hierarchically integrate ART, PV, and DL features across the input, feature, and decision levels. Experiments on the clinically collected multi-phase liver lesion (MPLL) dataset (the largest and most clinically comprehensive multi-phase liver cancer CECT dataset) demonstrate that the proposed method achieves state-of-the-art segmentation performance. `MADF-Net` achieves a Dice score of $\mathbf{78.65}\%$, which is $\mathbf{9.39}\%$ higher than the best single-stage baseline, by deeply fusing information from three phases, and consistently improves across all evaluation metrics. Our codes are available at https://anonymous.4open.science/r/ICLR26_unlocking_clinical_potential-EFE8/.

1 INTRODUCTION

Liver tumor segmentation is a critical task in quantitative medical image analysis, providing essential morphological and spatial information for surgical planning, radiotherapy, and post-treatment monitoring (Bilic et al., 2023). With the advent of deep learning, fully convolutional networks (FCNs), particularly U-Net and its variants (Ronneberger et al., 2015; Ren & Li, 2025; Du et al., 2022; Huang et al., 2017), have achieved notable success in automated segmentation tasks. These models extract features at either the 2D slice level or the 3D volumetric level, enabling robust representation learning for complex anatomical structures (Li et al., 2018).

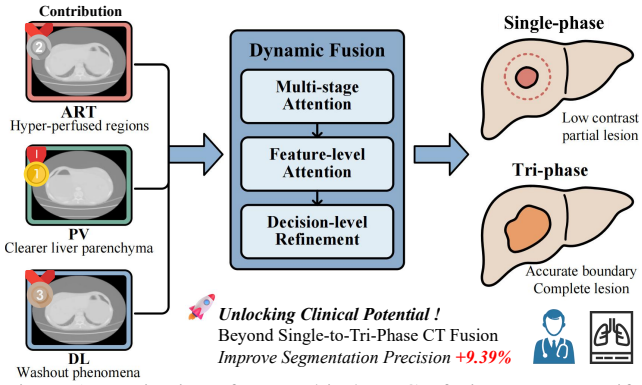

Figure 1: Motivation of our multi-phase CT fusion strategy. Different CT phases provide complementary information: ART highlights vessels, PV improves lesion–parenchyma contrast, and DL captures delayed enhancement. Their integration yields more accurate, robust liver tumor segmentation.

However, the majority of existing methods rely solely on single-phase Computed Tomography (CT) images (Bilic et al., 2023; Wang et al., 2023; Hatamizadeh et al., 2022; Cao et al., 2022), often ignoring the phase-specific characteristics inherent in clinical imaging protocols (Jun et al., 2023). Due to low tissue contrast and resolution limitations, single-phase methods frequently fail to achieve the precision required for clinical deployment

(Song et al., 2023; Zhang et al., 2024). Contrast-Enhanced CT (CECT), which captures dynamic changes in tissue attenuation following contrast agent administration, provides a valuable alternative by acquiring images at multiple time points—typically including the non-contrast (NC), arterial (ART), portal venous (PV), and delayed (DL) phases (Chi et al., 2013).

Among these, the NC phase offers a baseline anatomical information, but lacks enhancement patterns relevant to the tumor vasculature and lesion contrast, and is therefore generally not emphasized in liver tumor segmentation studies (Ni et al., 2024; Xu et al., 2021; Liu et al., 2024). The ART phase captures early vascular features, highlighting hyper-perfused regions and enhancing lesion boundary delineation (Kulkarni et al., 2021; Urban et al., 2000). The PV phase provides clearer liver parenchyma and structural completeness, facilitating more accurate segmentation (Kulkarni et al., 2021; Schneider et al., 2014). The DL phase captures delayed enhancement and washout phenomena, aiding in the identification of fibrotic or hypo-perfused tumors (Monzawa et al., 2007; Lim et al., 2002). The complementary nature of these phases offers a compelling opportunity for improved segmentation through multi-phase fusion (As shown in Figure 1). Therefore, how to effectively extract and fuse the features from different phases has attracted the attention of many researchers.

Existing multi-phase fusion strategies can be broadly categorized into three types (Zhang et al., 2021b): (1) *Input-level fusion* (Ouhmich et al., 2019), where multiple phases are concatenated as input and processed via a shared encoder; (2) *Feature-level fusion* (Zhang et al., 2021b; Zhu et al., 2022; Zhang et al., 2023; Hazirbas et al., 2016; Liu et al., 2023), which extracts features from each phase independently before combining them at intermediate layers; and (3) *Decision-level fusion* (Sun et al., 2017; Raju et al., 2020), where each phase is processed by a separate network and results are fused at the output level. While these approaches have shown potential, they often suffer from limitations such as insufficient modeling of nonlinear inter-phase relationships (Sun et al., 2017; Zhang et al., 2023), reduced reliability in ambiguous or low-contrast regions, and vulnerability to missing-phase scenarios common in clinical workflows (Xu et al., 2021; Zhu et al., 2022). Moreover, many existing methods treat all phases with equal importance during fusion, overlooking their distinct clinical value and the complementary information they offer (Xu et al., 2021; Zhong et al., 2024; Qiao et al., 2024). This results in suboptimal performance, especially in cases with blurred lesion boundaries or small lesions. **Therefore**, *how to effectively fuse multi-phase CT features while leveraging their individual strengths and mitigating their limitations remains an open challenge (Jiang et al., 2020).*

In this paper, we begin by systematically evaluating the segmentation performance of each enhanced CT phase using standard deep learning models. Our quantitative analysis reveals that the PV phase contributes most significantly to segmentation accuracy, consistent with its known clinical role. Guided by this observation, we propose a novel framework, Multi-phase Attention Deep Fusion Network (`MADF-Net`), to exploit the complementary advantages of the ART, PV, and DL phases through hierarchical fusion. `MADF-Net` introduces full-stage attention-based fusion across the input, feature, and decision levels, enabling deep inter-phase information interaction. Extensive experiments on the MPLL dataset demonstrate that our method achieves state-of-the-art segmentation performance, reaching a Dice score of 78.65% when using all three phases, representing a 9.39% improvement over the best single-phase baseline, confirming its robustness and generalizability.

Our contributions are as follows:

- ❶ We conduct a comprehensive quantitative analysis of liver tumor segmentation across different CT phases and demonstrate the predominant contribution of the PV phase, providing both empirical and clinical insights.
- ❷ We propose `MADF-Net`, a multi-phase attention-based fusion network that integrates ART, PV, and DL phase features at multiple stages, enhancing liver tumor segmentation performance through deep inter-phase feature interaction.
- ❸ Extensive experiments on a newly collected multi-phase liver lesion (MPLL) benchmark (the largest and most clinically comprehensive multi-phase liver cancer CECT dataset) demonstrate that the proposed method achieves state-of-the-art liver tumor segmentation performance.

2 RELATED WORKS

Single-Phase Based Liver Tumor Segmentation. Deep learning has significantly advanced single-phase liver tumor segmentation in CT images. Ronneberger et al. (Ronneberger et al., 2015)

introduced U-Net, whose encoder-decoder structure with skip connections became foundational, effectively capturing both local details and global context for handling low contrast and fuzzy boundaries. H-DenseUNet (Li et al., 2018) enhanced feature reuse through hybrid dense connections and achieved state-of-the-art results on the LiTS2017 dataset. Variants such as UNet++ (Zhou et al., 2018) further improved efficiency and multi-scale accuracy, particularly for small tumor detection. However, CNN-based models often struggled with global context in complex tumor structures. To address this, TransUNet (Chen et al., 2021) combined CNNs for low-level feature extraction with Transformers for global dependency modeling. UNETR (Hatamizadeh et al., 2022) and UNETR++ (Shaker et al., 2024) integrated global context and local detail via a Transformer-based U-shaped architecture, achieving strong performance on 3D CT tasks. Nevertheless, single-phase methods still suffer from limited tissue contrast and resolution, resulting in information loss and reduced clinical applicability.

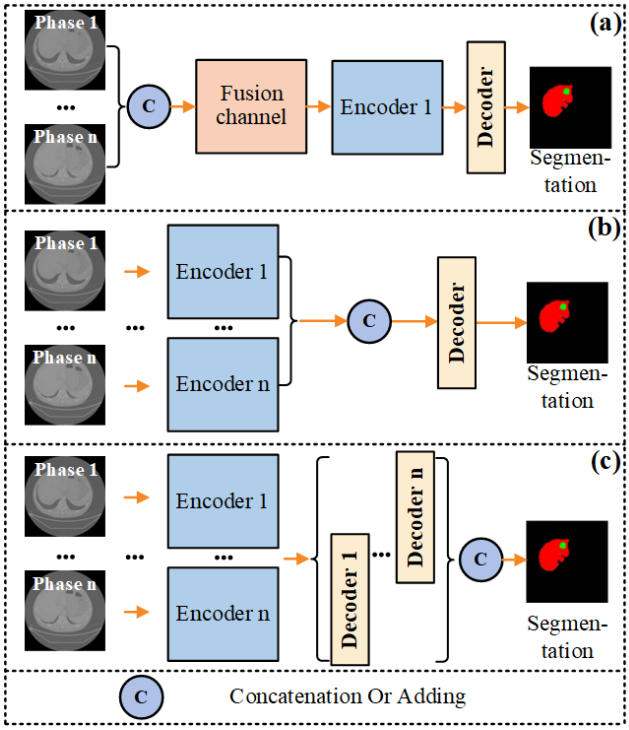

Figure 2: Multi-phase fusion method of enhanced CT. (a), (b), and (c) correspond to input-level, feature-level, and decision-level fusion architectures, respectively.

Multi-Phase Based Liver Tumor Segmentation. Recently, an increasing number of studies have investigated how to leverage multi-phase CT information to improve liver tumor segmentation performance. Multiphase fusion is typically performed at one of three stages: input-level, feature-level, or decision-level fusion (Zhang et al., 2021b), referred to as single-stage fusion in this paper. Alternatively, fusion can occur across multiple stages, which we define as multi-stage fusion. ❶ **Single-stage Fusion.** An early example of single-stage input-level fusion was proposed by Ouhmich et al. (Ouhmich et al., 2019), who concatenated PV and ART phase images as input to a U-Net, significantly improving tumor segmentation performance over single-phase training. Feature-level fusion is currently the most active research area. Zhou et al. (Zhou et al., 2019) introduced a dual-path 3D fully convolutional network with cross-phase skip connections to enable dense information exchange. Wu et al. (Wu et al., 2019) treated non-contrast and enhanced CT scans equally and applied feature-level fusion at selected U-Net layers. In decision-level fusion, features are independently extracted from each phase and fused at higher layers. Raju et al. (Raju et al., 2020) proposed an integrated joint and semi-supervised training strategy that leveraged limited plain and enhanced CT data to achieve robust cross-domain segmentation. Despite progress in single-stage fusion, these methods still face challenges such as information loss and limited ability to capture phase-specific characteristics. ❷ **Multi-stage Fusion.** Feature-level and decision-level combinations currently dominate multi-stage fusion network designs (Ni et al., 2024; Liu et al., 2024; Zhu et al., 2022; Kuang et al., 2024). PA-ResSeg (Xu et al., 2021) introduced intra- and inter-phase attention mechanisms to capture both channel-wise dependencies and cross-phase interactions, embedding attentional modules at each encoder layer to fuse multi-scale information from ART and PV phases. Building on this, SA-Net (Zhang et al., 2021b) added a spatial aggregation module for encoding-stage interaction and an uncertainty correction module at the decision stage to refine fuzzy tumor boundaries. To address spatial misalignment in multi-phase CT, Zhang et al. (Zhang et al., 2023) incorporated differentiable deformation operations (Jaderberg et al., 2015) for enhanced feature alignment. Raju et al. (Raju et al., 2020) proposed a joint and semi-supervised training strategy that effectively leveraged limited non-contrast and enhanced CT data, though at the cost of increased training time. HRadNet (Liang et al., 2023) utilized a feature

pyramid and a metadata fusion layer to incorporate clinical features such as tumor size and patient age, improving generalizability. However, most existing multi-stage approaches adopt only two fusion stages and still suffer from potential information loss, limiting segmentation accuracy and robustness. To address this, we propose a three-stage fusion network designed to better preserve information throughout the extraction and fusion process.

3 PRELIMINARY

To fully exploit the complementary anatomical and pathological information provided by the three phases of CECT, we propose a novel three-stage fusion framework, named Multi-phase Attention Deep Fusion Network (`MADF-Net`). This section first introduces preliminary knowledge on fusion strategies, and then describes the proposed `MADF-Net`. As shown in Figure 2, three common fusion strategies are illustrated. A detailed pseudocode description of the overall procedure is provided in Appendix A.

Input-level Fusion. As shown in Figure 2 (a), this strategy concatenates images from different phases (Phase 1, Phase 2, ..., Phase n) along the channel dimension at the input stage to form a unified input tensor. To enhance flexibility, we introduce learnable phase-wise modulation weights $\{\alpha_i\}_{i=1}^n$ and a normalization operator $\mathcal{N}(\cdot)$:

$$\boldsymbol{I}_{\text{input}} = \mathcal{N}\left(\Big\|_{i=1}^{n} \left(\alpha_i \cdot \Gamma(\boldsymbol{I}_i) + \beta_i \cdot \mathbf{1}_{H\times W\times C}\right) \right), \alpha_i = \frac{\exp(\theta_i)}{\sum_{j=1}^{n}\exp(\theta_j)}, \tag{1}$$

where $\Gamma(\cdot)$ denotes intensity standardization, $\boldsymbol{I}_i \in \mathbb{R}^{H\times W\times C}$ is the i-th phase image, θ_i are learnable logits and $\|$ denotes channel-wise concatenation. This formulation adaptively highlights more informative phases while suppressing noisy ones.

Feature-level Fusion. As shown in Figure 2 (b), this strategy integrates features from multiple phases by using attention-guided gating and nonlinear projections. Let $\boldsymbol{E}_i \in \mathbb{R}^{H'\times W'\times d}$ be the feature map extracted from the i-th phase. We compute phase attention maps $\boldsymbol{A}_i$ from global descriptors $\boldsymbol{g}_i$ via a softmax-normalized MLP, and then fuse features as:

$$\boldsymbol{E}_{\text{fusion}} = \Psi\left(\sum_{i=1}^{n} \Bigg(\underbrace{\frac{\exp(\mathbf{W}_a \boldsymbol{g}_i + \mathbf{b}_a)}{\sum_{j=1}^{n}\exp(\mathbf{W}_a \boldsymbol{g}_j + \mathbf{b}_a)}}_{\boldsymbol{A}_i} \odot \underbrace{\left(\mathbf{W}_v * \boldsymbol{E}_i + \mathbf{b}_v\right)}_{\tilde{\boldsymbol{E}}_i} \Bigg) \right), \boldsymbol{g}_i = \text{GAP}(\boldsymbol{E}_i), \tag{2}$$

where $*$ denotes convolution, $\text{GAP}(\cdot)$ is global average pooling, and $\Psi(\cdot)$ is a residual refinement block. This design allows adaptive semantic fusion guided by global context cues.

Decision-level Fusion. As shown in Figure 2 (c), this strategy constructs separate segmentation heads for each phase, and aggregates the resulting predictions $\{\boldsymbol{S}_i\}_{i=1}^n$ based on their confidence. We employ an uncertainty-aware soft weighting scheme with temperature scaling:

$$\boldsymbol{S}_{\text{final}} = \sum_{i=1}^{n} \underbrace{\left(\frac{\exp\left(-\tau \cdot \mathcal{H}(\boldsymbol{S}_i)\right)}{\sum_{j=1}^{n}\exp\left(-\tau \cdot \mathcal{H}(\boldsymbol{S}_j)\right)} \right)}_{w_i} \cdot \underbrace{\sigma(\boldsymbol{S}_i)}_{\text{probability map}}, \mathcal{H}(\boldsymbol{S}_i) = -\frac{1}{|\Omega|}\sum_{p\in\Omega}\sum_{c} \boldsymbol{S}_i^{(p,c)} \log \boldsymbol{S}_i^{(p,c)}, \tag{3}$$

where $\mathcal{H}(\cdot)$ computes the spatial entropy over pixel set Ω, and τ controls weight sharpness. This formulation emphasizes confident predictions and suppresses noisy ones for decision-level fusion.

4 METHODOLOGY

To address the challenge of liver tumor segmentation using multiphase CT data, we propose a unified framework, `MADF-Net`, that integrates input-level, feature-level, and decision-level fusion. As shown in Figure 3, the network consists of two parallel branches (main and auxiliary) with symmetric encoder-decoder structures, enabling hierarchical feature aggregation across three CT phase.

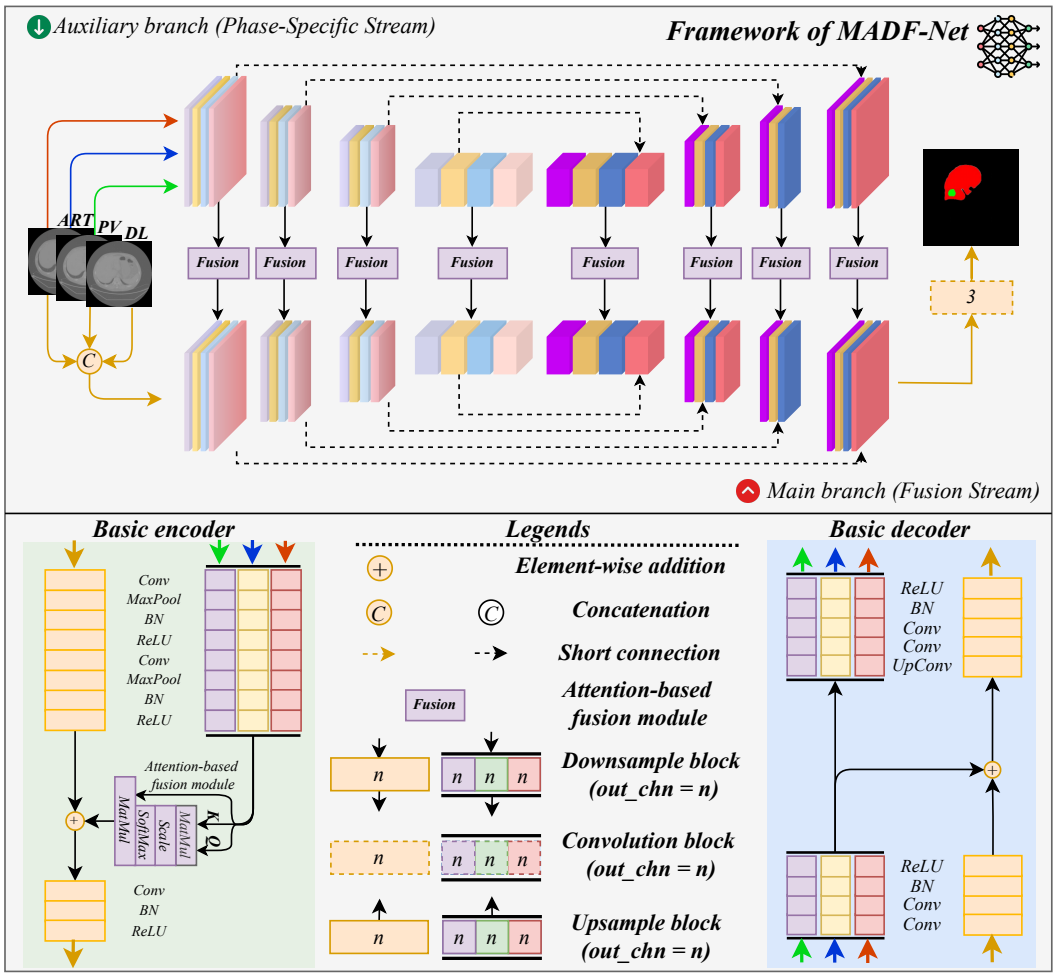

Figure 3: Overview of MADF-Net. Three single-phase CT inputs (ART, PV, DL) are processed via parallel main/auxiliary branches. Fusion occurs at three levels: (1) Input-level (concatenation of phases in main branch); (2) Feature-level (cross-branch fusion in encoder/decoder blocks); (3) Decision-level (final result aggregation).

Input-Level Fusion: Multi-Phase Data Initialization. Given the three CT phases $\boldsymbol{I}_{ART}$, $\boldsymbol{I}_{PV}$, and $\boldsymbol{I}_{DL}$, the main branch performs early-stage fusion by channel-wise concatenation to directly expose the encoder to cross-phase correlations: $\boldsymbol{I}_{fusion}$= Concat$(\boldsymbol{I}_{ART}, \boldsymbol{I}_{PV}, \boldsymbol{I}_{DL}) \in \mathbb{R}^{H\times W\times 3C}$, where H and W denote the spatial resolution and C denotes the number of channels per phase. In parallel, the auxiliary branch independently forwards each phase through three isolated sub-encoders: $\boldsymbol{I}^{(p)}_{phase}$= $\{\boldsymbol{I}_{ART}, \boldsymbol{I}_{PV}, \boldsymbol{I}_{DL}\}$, $p \in \{1, 2, 3\}$, preserving phase-specific characteristics that might otherwise be suppressed by early fusion.

Encoder Blocks: Hierarchical Feature Extraction with Cross-Branch Fusion. Both branches comprise four encoder blocks indexed by $l \in \{1, \ldots, 4\}$, progressively downsample spatial resolution while expanding feature dimensionality (e.g., $C \to 4C \to 16C \to 64C$). Encoder block performs following operations: **(I) Self-Attention-based Phase Reweighting.** For auxiliary features $\boldsymbol{X}^{(l)}_{ART}$, $\boldsymbol{X}^{(l)}_{PV}$, and $\boldsymbol{X}^{(l)}_{DL}$, a shared self-attention module computes attention weights across phases:

$$\boldsymbol{\alpha}^{(l)} = \text{softmax}\left(\frac{\boldsymbol{Q}^{(l)} \cdot \left(\boldsymbol{K}^{(l)}\right)^{\top}}{\sqrt{d_k}}\right), \quad \boldsymbol{Q}^{(l)} = \boldsymbol{W}_Q \boldsymbol{X}^{(l)}, \; \boldsymbol{K}^{(l)} = \boldsymbol{W}_K \boldsymbol{X}^{(l)}, \tag{4}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_k \times d_{in}}$ are learnable projections and $\boldsymbol{X}^{(l)} = \left[\boldsymbol{X}^{(l)}_{ART}, \boldsymbol{X}^{(l)}_{PV}, \boldsymbol{X}^{(l)}_{DL}\right]$. The reweighted auxiliary feature is then: $\boldsymbol{X}^{(l)}_{aux}$= $\sum \alpha^{(l)}_p \odot \boldsymbol{X}^{(l)}_p$, $p \in \{ART, PV, DL\}$

(II) Feature-level Cross-Branch Fusion. After obtaining the reweighted auxiliary representation, we further integrate it with the main branch feature $\boldsymbol{X}_{main}^{(l)}$. To this end, we design a *gated residual summation* mechanism:

$$\boldsymbol{X}_{enc}^{(l)} = \sigma\Big(\boldsymbol{W}_g * [\boldsymbol{X}_{main}^{(l)}, \boldsymbol{X}_{aux}^{(l)}]\Big) \odot \boldsymbol{X}_{main}^{(l)} + \Big(1 - \sigma\Big(\boldsymbol{W}_g * [\boldsymbol{X}_{main}^{(l)}, \boldsymbol{X}_{aux}^{(l)}]\Big)\Big) \odot \boldsymbol{X}_{aux}^{(l)}, \quad (5)$$

where the gating factor is adaptively determined by the concatenated representations from both branches. This allows the model to dynamically balance their contributions, avoiding redundancy and gradient dilution caused by naive concatenation or summation. Specifically, when the gate approaches 1, the model emphasizes high-level semantics from the main branch, while values closer to 0 highlight fine-grained cues from the auxiliary branch.

Nevertheless, simply stacking the above fusion operation across multiple depths may hinder gradient propagation, thus limiting the representation capacity of deeper layers. To address this, we introduce a *residual preservation regularization* to facilitate cross-layer information flow. Concretely, at the l-th layer, the fused feature $\boldsymbol{X}_{enc}^{(l)}$ is enhanced with a gated residual connection from previous layer:

$$\hat{\boldsymbol{X}}_{enc}^{(l)} = \boldsymbol{X}_{enc}^{(l)} + \lambda \cdot \Big(\alpha^{(l)} \odot \hat{\boldsymbol{X}}_{enc}^{(l-1)} + \big(1 - \alpha^{(l)}\big) \odot f\Big(\hat{\boldsymbol{X}}_{enc}^{(l-1)}\Big)\Big), \quad l > 1, \quad (6)$$

where λ is a learnable global balancing coefficient, and $\alpha^{(l)}$ is a layer-wise gating vector that adaptively controls the trade-off between direct residual propagation and a transformed path. The function $f(\cdot)$ denotes a lightweight non-linear mapping (e.g., a convolutional projection or an MLP). This design ensures that shallow features can effectively penetrate deeper layers to improve gradient flow, while the nonlinear transformation path enriches cross-layer feature diversity.

Decoder Blocks: Multi-Scale Feature Reconstruction. The decoder consists of four blocks that mirror the encoder structure, progressively upsampling the fused representations $\boldsymbol{X}_{enc}^{(l)}$ back to the original resolution. Each decoder block not only restores spatial resolution but also selectively incorporates complementary information from shallow layers through gated skip connections. Specifically, at each stage l, we compute:

$$\boldsymbol{Y}^{(l-1)} = \mathrm{ReLU}\Big(\boldsymbol{G}^{(l)} \odot \mathrm{UpConv}\big(\boldsymbol{X}_{enc}^{(l)}\big) + \big(1 - \boldsymbol{G}^{(l)}\big) \odot \mathcal{F}_{att}\big(\boldsymbol{X}_{enc}^{(l-1)}, \mathrm{UpConv}(\boldsymbol{X}_{enc}^{(l)})\big)\Big), \quad (7)$$

where $\boldsymbol{G}^{(l)}$ is a learned gating map, $\mathrm{UpConv}(\cdot)$ denotes an upsampling convolution block with batch normalization and activation, and $\mathcal{F}_{att}(\cdot,\cdot)$ is an attention-based fusion module for shallow-deep interaction. This stage-wise design ensures that high-resolution details from earlier layers are progressively blended with the deep semantic context from later layers.

To further enhance multi-scale consistency and stabilize gradient flow, we augment the reconstruction with a residual-preserving multi-scale aggregation term:

$$\hat{\boldsymbol{Y}}^{(l-1)} = \boldsymbol{Y}^{(l-1)} + \mu \cdot \Big(\sum_{k=1}^{K} \beta_k^{(l)} \odot \mathrm{UpConv}^{(k)}\Big(\boldsymbol{X}_{enc}^{(l-k)}\Big)\Big), \quad (8)$$

where μ is a learnable global scaling factor, $\beta_k^{(l)}$ are adaptive weights normalized by a softmax constraint, and $\mathrm{UpConv}^{(k)}(\cdot)$ denotes k-step hierarchical upsampling operators. This formulation explicitly aggregates contextual evidence from multiple encoder depths, enabling the decoder to reconstruct fine details while preserving long-range semantic dependencies.

In summary, the decoder leverages a combination of gated skip fusion, residual-preserving connections, and multi-scale aggregation to ensure both spatial fidelity and semantic consistency. Such a design alleviates the common issue of blurred boundaries in dense prediction tasks, while also enhancing robustness against vanishing gradients during backpropagation.

Decision-Level Fusion: Final Segmentation Output. Finally, the outputs of the main and auxiliary decoders are aggregated to produce the segmentation mask. Specifically,

$$\boldsymbol{O}_{final} = \sigma\Big(\boldsymbol{W}_{out} * \big[\boldsymbol{O}_{ART}, \boldsymbol{O}_{PV}, \boldsymbol{O}_{DL}, \boldsymbol{O}_{fusion}\big] + \boldsymbol{b}_{out}\Big), \quad (9)$$

where $\boldsymbol{O}_{ART}, \boldsymbol{O}_{PV}, \boldsymbol{O}_{DL}$ are the three auxiliary outputs, $\boldsymbol{O}_{fusion}$ is the main-branch output, and $\boldsymbol{W}_{out}, \boldsymbol{b}_{out}$ denote the parameters of the final convolutional projection. This decision-level fusion enforces complementary exploitation of both phase-specific and cross-phase knowledge, yielding a precise tumor segmentation mask.

5 EXPERIMENTS

Dataset Curation (Multi-phase Dataset). The multi-phase liver lesion (MPLL) dataset, consists of **952, 601** 2D slices with liver disease from the "Anonymous Authoritative Hospitals (Information will be made public after the paper is accepted)". The dataset includes patients aged between 9 and 72 years, and the number of axial slices per scan varying from 48 to 777. This is the largest and most valuable multi-phase CECT liver cancer dataset to date, all the images in MPLL dataset contain three enhanced phases (ART, PV, and DL). The registered images were annotated using ITK-SNAP software by two experienced attending radiologists, and subsequently reviewed by a third attending radiologist to ensure the accuracy and consistency of the annotations. All data have been anonymized and contain only image information. The MPLL dataset has received approval from the institutional ethics committee under certification number 2022-BE(H)-194. Figure 6 shows example images from the datasets. The training, validation, and test splits (7:1:2, following previous work (Jiang et al., 2023)), along with image dimensions and other details, are summarized in Table 1. A more detailed description is provided in Appendix B.

Table 1: **Dataset characteristics.**

Dataset	Attribute	Value
MPLL	Phase	ART, PV, DL
	Slice thickness	0.62mm–5.0mm
	Slice resolution	512 × 512
	Disease type	ABS, HCC, HEM, ICC, Lipoma

Evaluation Metrics. We employed the Dice Similarity Coefficient (DSC), Jaccard Similarity Coefficient (JSC), Average Symmetric Surface Distance (ASSD), and 95% Hausdorff Distance (HD_{95}) (Jiang et al., 2025) to evaluate the experimental results. In the experiments on single-phase (1P), two-phase (2P) and three-phase (3P) input, we additionally employed Volume Overlap Error (VOE) and Relative Volume Difference (RVD) as supplementary evaluation metrics.

Implementation Details. All models were trained for 100 epochs with a batch size of 8. The Stochastic Gradient Descent (SGD) optimizer was adopted with a learning rate of 0.01 and 4 parallel data loading workers. Data augmentation techniques include horizontal flipping (with probability 0.5) and vertical flipping (with probability 0.5), and no post-processing is used.

The proposed method was implemented on a Linux 5.4.0 system using PyTorch 1.13.1. All experiments were conducted on two NVIDIA GeForce RTX 3090 GPUs (24 GB × 2), providing sufficient computational resources for efficient model training and evaluation.

5.1 MAIN RESULT

Table 2: Quantitative comparison of segmentation performance across different phase combinations (relative to 3P, based on numerical differences).

Input Phases	DSC(%)↑	JSC(%)↑	HD_{95}↓	ASSD↓
1P (PV)	$69.26_{\downarrow 9.39}$	$64.08_{\downarrow 10.73}$	$40.298_{\uparrow 13.492}$	$16.958_{\uparrow 6.222}$
2P (ART+PV)	$76.09_{\downarrow 2.56}$	$71.41_{\downarrow 3.40}$	$28.838_{\uparrow 2.032}$	$15.659_{\uparrow 4.923}$
3P (ART+PV+DL)	**78.65**	**74.81**	**26.806**	**10.736**

Obs. ❶: Phase Combination Analysis. We conducted three groups of experiments to explore the optimal input combination for 1P, 2P, and 3P settings. As listed in Table 2, the number of input phases increases, segmentation accuracy improves accordingly, demonstrating the effectiveness of the proposed `MADF-Net` in leveraging complementary information across multiple imaging phases. Specifically, the best result of 1P input (Zheng et al., 2024b) using only the PV phase achieves a DSC of 69.26% and a JSC of 64.08%. When the ART phase is added to form the 2P input (ART+PV), both DSC and JSC show moderate improvements to 76.09% and 71.41%, respectively, while the HD_{95} drops significantly from 40.298 to 28.838, indicating better boundary localization. The 3P input (ART+PV+DL) achieves the best overall performance, with the highest DSC (78.65%) and JSC (74.81%), along with the lowest HD_{95} (26.80) and Average ASSD of 10.73. These results suggest that the additional information from the ART and DL phases enhances both global overlap and local boundary accuracy. Compared to the 1P setting, the 3P input yields substantial gains of 9.39% in DSC and 10.73% in JSC, underscoring the benefit of multi-phase fusion in capturing diverse tumor characteristics.

Obs. ❷: Comparison of Single Phase Performance. We conducted a segmentation performance comparison using different single-phase input images on the MPLL dataset, evaluating several state-of-the-art models, including FANet (Tomar et al., 2022), GRENet (Wang et al., 2023), ASSNet (Zheng et al., 2024a), TransUNet (Chen et al., 2021), KiU-Net (Valanarasu et al., 2021),

Table 3: Quantitative comparison of different methods on the MPLL dataset. Best results are **bold**, and differences relative to PV are shown as colored arrows.

Methods	DSC(%)↑			Jaccard(%)↑			HD_{95} (mm)↓			ASSD (mm)↓		
	ART	PV	DL	ART	PV	DL	ART	PV	DL	ART	PV	DL
FANet	$67.56_{\downarrow 1.52}$	**69.08**	$66.01_{\downarrow 3.07}$	$61.98_{\downarrow 1.11}$	**63.09**	$60.74_{\downarrow 2.35}$	$60.384_{\uparrow 17.529}$	**42.855**	$62.083_{\uparrow 19.228}$	$30.300_{\uparrow 10.107}$	**20.193**	$25.206_{\uparrow 5.013}$
GRENet	$66.59_{\downarrow 1.67}$	**68.26**	$66.21_{\downarrow 2.05}$	$60.14_{\downarrow 0.94}$	**61.08**	$59.91_{\downarrow 1.17}$	$60.651_{\uparrow 3.390}$	**57.261**	$65.412_{\uparrow 8.151}$	$28.069_{\uparrow 4.059}$	**24.010**	$30.982_{\uparrow 6.972}$
ASSNet	$67.91_{\downarrow 1.35}$	**69.26**	$66.49_{\downarrow 2.77}$	$60.01_{\downarrow 4.07}$	**64.08**	$60.07_{\downarrow 4.01}$	$45.260_{\uparrow 4.962}$	**40.298**	$59.946_{\uparrow 19.648}$	$22.074_{\uparrow 5.116}$	**16.958**	$25.031_{\uparrow 8.073}$
TransUNet	$67.51_{\downarrow 0.97}$	**68.48**	$67.15_{\downarrow 1.33}$	$58.65_{\downarrow 2.58}$	**61.23**	$60.53_{\downarrow 0.70}$	$55.960_{\downarrow -0.339}$	**55.301**	$58.594_{\uparrow 3.293}$	$26.983_{\downarrow 0.346}$	**27.329**	$24.972_{\downarrow 2.357}$
KiU-Net	$65.92_{\downarrow 1.70}$	**67.62**	$65.62_{\downarrow 2.00}$	$59.64_{\downarrow 1.24}$	**60.88**	$59.42_{\downarrow 1.46}$	$62.902_{\uparrow 3.642}$	**59.260**	$65.956_{\uparrow 6.696}$	$30.821_{\uparrow 2.178}$	**28.643**	$27.983_{\downarrow 0.660}$
AttUNet	$66.01_{\downarrow 2.51}$	**68.52**	$65.85_{\downarrow 2.67}$	$59.82_{\downarrow 1.84}$	**61.66**	$59.61_{\downarrow 2.05}$	$60.913_{\uparrow 5.289}$	**55.624**	$66.583_{\uparrow 10.959}$	$31.098_{\uparrow 4.117}$	**26.981**	$27.973_{\downarrow -3.008}$

and AttUNet (Chen et al., 2023). The results are summarized in Table 3. Across all models, segmentation performance was consistently better on the PV phase compared to the ART and DL phases. For example, with FANet, the DSC and JSC on PV were 1.52% and 1.11% higher than on ART, and 3.07% and 2.35% higher than on DL, respectively. In terms of boundary metrics, the HD_{95} and ASSD on PV were 17.529 and 10.107 lower than on ART, and 19.228 and 5.013 lower than on DL, indicating more accurate boundary localization. These results suggest that segmentation outputs on the PV phase more closely match the ground truth. In clinical contexts, this may be attributed to the PV phase offering more distinct grayscale contrast between tissues, as well as between lesions and normal structures (Ni et al., 2024; Lam et al., 2017; Al-Battal et al., 2024). This contrast enhancement is particularly beneficial in cases with ambiguous or highly heterogeneous tumor boundaries (Liu et al., 2024).

Obs. ❸: 2P Fusion: ART and PV. Single-phase experimental results indicate that segmentation performance is high when using the PV and ART phases as inputs. Based on this observation, we conducted 2P (A+P) experiments on the MPLL dataset, comparing the proposed `MADF-Net` with several state-of-the-art multi-phase segmentation methods, including MAML (Zhang et al., 2021a), MW-UNet (Zhu et al., 2022), SA-Net (Zhang et al., 2021b), PA-ResSeg (Xu et al., 2021), and MCDA-Net (Kuang et al., 2024). As listed in Table 4, `MADF-Net` achieves superior performance, improving the DSC metric by 5.27%, 4.71%, 4.18%, 1.90%, and 0.01% over the five comparison methods, respectively. It also consistently outperforms all other methods in terms of HD_{95}, demonstrating its ability to exploit complementary information across imaging phases to enhance both segmentation accuracy and boundary localization.

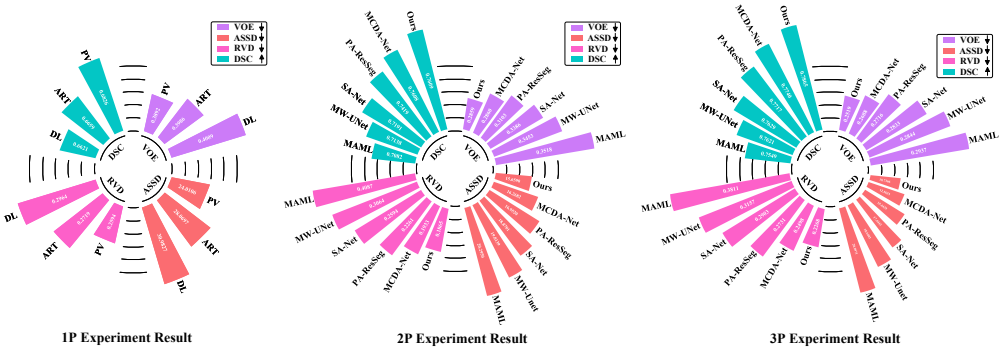

Figure 4: Comparison of the segmentation performance on 1P, 2P and 3P, evaluated using multiple quantitative metrics including DSC, VOE, RVD, and ASSD. Specifically, the black upward arrow (↑) denotes that a higher metric value indicates superior performance, while the black downward arrow (↓) signifies that a lower metric value reflects more favorable outcomes.

Table 4: Quantitative comparison of 2-phase (left) and 3-phase (right) inputs. "Phase" indicates the data modality used. The best are highlighted in **bold**.

Method	2-Phase (A+P)				3-Phase (A+P+D)			
	DSC(%)↑	JSC(%)↑	HD_{95} (mm)↓	ASSD (mm)↓	DSC(%)↑	JSC(%)↑	HD_{95} (mm)↓	ASSD (mm)↓
MAML	$70.82_{\downarrow 5.27}$	$64.82_{\downarrow 6.59}$	$53.223_{\uparrow 24.38}$	$20.295_{\uparrow 4.64}$	$75.49_{\downarrow 3.16}$	$70.63_{\downarrow 4.18}$	$41.588_{\uparrow 14.78}$	$20.097_{\uparrow 9.36}$
MW-UNet	$71.38_{\downarrow 4.71}$	$65.47_{\downarrow 5.94}$	$37.519_{\uparrow 8.68}$	$19.013_{\uparrow 3.35}$	$76.21_{\downarrow 2.44}$	$71.56_{\downarrow 3.25}$	$35.221_{\uparrow 8.42}$	$18.640_{\uparrow 7.90}$
SA-Net	$71.91_{\downarrow 4.18}$	$66.14_{\downarrow 5.27}$	$36.946_{\uparrow 8.11}$	$18.870_{\uparrow 3.21}$	$76.29_{\downarrow 2.36}$	$71.67_{\downarrow 3.14}$	$34.126_{\uparrow 7.32}$	$17.444_{\uparrow 6.71}$
PA-ResSeg	$74.19_{\downarrow 1.90}$	$68.97_{\downarrow 2.44}$	$31.287_{\uparrow 2.45}$	$16.952_{\uparrow 1.29}$	$77.17_{\downarrow 1.48}$	$72.84_{\downarrow 1.97}$	$33.608_{\uparrow 6.80}$	$15.443_{\uparrow 4.71}$
MCDA-Net	$76.08_{\downarrow 0.01}$	$71.40_{\downarrow 0.01}$	$30.436_{\uparrow 1.60}$	$16.268_{\uparrow 0.61}$	$77.40_{\downarrow 1.25}$	$73.12_{\downarrow 1.69}$	$28.600_{\uparrow 1.79}$	$12.565_{\uparrow 1.83}$
`MADF-Net`	**76.09**	**71.41**	**28.8382**	**15.6590**	**78.65**	**74.81**	**26.8068**	**10.7366**

Obs. ❹: 3P Fusion: ART, PV, and DL. As listed in Table 4, the proposed `MADF-Net` achieved the state-of-the-art performance in terms of DSC (78.65%), JSC (74.81%), HD_{95} (26.806), and ASSD (10.736) compared to the other five methods on the 3P fusion strategy. This indicates that our `MADF-Net` more accurately localizes the spatial positions and delineates the geometric shapes of the target regions. The quantitative comparison of the performance across 1P, 2P, and 3P fusion strategies is further illustrated in Figure 4, and the visualization of the segmentation results is shown in Figure 5. Compared to other methods, the proposed `MADF-Net` achieves the closest performance to the ground truth in terms of tumor contour localization and small-object boundary segmentation.

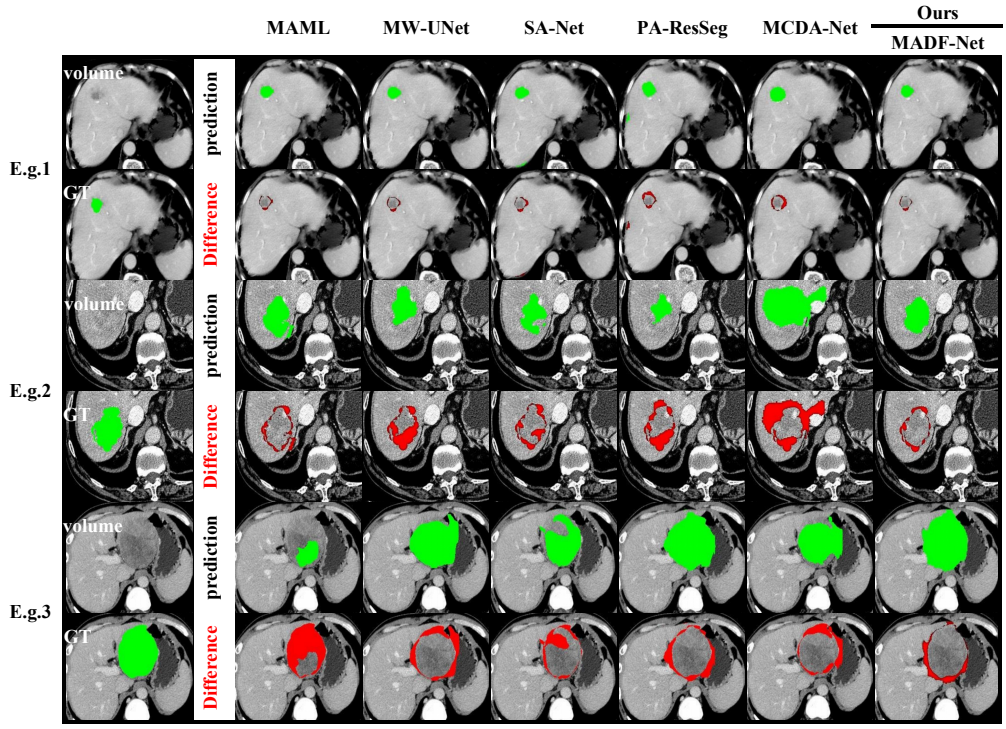

Figure 5: Result comparison of different three-phase networks. For better visualization, we performed appropriate cropping. The green region in the ground truth (GT) row represents the tumor, the green region in the prediction row indicates the predicted tumor area, and the red region in the difference row denotes the difference between the two.

5.2 Efficiency Analysis

Table 5 compares `MADF-Net` with baseline models. Our model requires 99.552×10^9 GFLOPs and 40.482 M parameters, achieving a favorable balance between computational cost and model size compared to SA-Net (152.965×10^9 GFLOPs, 170.852 M) and PA-ResSeg (64.660×10^9 GFLOPs, 67.732 M), while remaining competitive with lighter models such as MAML and MW-UNet. Further details on the experimental setup and efficiency comparisons are provided in Appendix C.

6 Conclusion and Future Work

This paper presented `MADF-Net`, a novel multi-phase attention-based fusion network for liver tumor segmentation in contrast-enhanced CT images. Our approach is guided by a systematic quantitative evaluation of individual phases, confirming the predominant contribution of the PV phase and its alignment with clinical understanding. `MADF-Net` performs full-stage fusion across the input, feature, and decision levels to fully exploit the complementary information from ART, PV, and DL phases. The experiment on MPLL datasets demonstrate that our method achieves state-of-the-art performance and generalizes well across datasets. **Future Work:** We will (i) design phase-specific subnetworks and investigate phase-aware pretraining, and (ii) extend the paradigm to multi-modal/multi-omics fusion by integrating CT with digital pathology and radiomics-derived omics for comprehensive patient-level modeling.

REFERENCES

Abdullah F Al-Battal, Soan Duong, Van Ha Tang, Quang Duc Tran, Steven QH Truong, Chien Phan, Truong Q Nguyen, and Cheolhong An. Multi-target and multi-stage liver lesion segmentation and detection in multi-phase computed tomography scans. *arXiv preprint arXiv:2404.11152*, 2024.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pp. 205–218. Springer, 2022.

Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):55–68, 2023.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Yanling Chi, Jiayin Zhou, Sudhakar K Venkatesh, Qi Tian, and Jimin Liu. Content-based image retrieval of multiphase ct images for focal liver lesion characterization. *Medical Physics*, 40(10): 103502, 2013.

Hao Du, Jiazheng Wang, Min Liu, Yaonan Wang, and Erik Meijering. Swinpa-net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5355–5366, 2022.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pp. 213–228. Springer, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4700–4708, 2017.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Chunhui Jiang, Yi Wang, Qingni Yuan, Pengju Qu, and Heng Li. A 3d medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. *Scientific Reports*, 15(1):6159, 2025.

Linfeng Jiang, Jiajie Ou, Ruihua Liu, Yangyang Zou, Ting Xie, Hanguang Xiao, and Ting Bai. Rmau-net: Residual multi-scale attention u-net for liver and tumor segmentation in ct images. *Computers in Biology and Medicine*, 158:106838, 2023.

Xixi Jiang, Qingqing Luo, Zhiwei Wang, Tao Mei, Yu Wen, Xin Li, Kwang-Ting Cheng, and Xin Yang. Multi-phase and multi-level selective feature fusion for automated pancreas segmentation from ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 460–469. Springer, 2020.

Eunji Jun, Seungwoo Jeong, Da-Woon Heo, and Heung-Il Suk. Medical transformer: Universal encoder for 3-d brain mri analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Haopeng Kuang, Xue Yang, Hongjun Li, Jingwei Wei, and Lihua Zhang. Adaptive multiphase liver tumor segmentation with multiscale supervision. *IEEE Signal Processing Letters*, 31:426–430, 2024.

Naveen Kulkarni, Alice Fung, Avinash R Kambadakone, and Benjamin M Yeh. Ct techniques, protocols, advancements and future directions in liver diseases. *Magnetic Resonance Imaging Clinics of North America*, 29(3):305, 2021.

A Lam, D Fernando, CC Sirlin, M Nayyar, SC Goodwin, DK Imagawa, and C Lall. Value of the portal venous phase in evaluation of treated hepatocellular carcinoma following transcatheter arterial chemoembolisation. *Clinical Radiology*, 72(11):994–e9, 2017.

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

Yinhao Liang, Wenjie Tang, Ting Wang, Wing WY Ng, Siyi Chen, Kuiming Jiang, Xinhua Wei, Xinqing Jiang, and Yuan Guo. Hradnet: A hierarchical radiomics-based network for multicenter breast cancer molecular subtypes prediction. *IEEE Transactions on Medical Imaging*, 43(3): 1225–1236, 2023.

Jae Hoon Lim, Dongil Choi, Seung Hoon Kim, Soon Jin Lee, Won Jae Lee, Hyo Keun Lim, and Seonwoo Kim. Detection of hepatocellular carcinoma: value of adding delayed phase imaging to dual-phase helical ct. *American Journal of Roentgenology*, 179(1):67–73, 2002.

Lihao Liu, Angelica I Aviles-Rivero, and Carola-Bibiane Schönlieb. Contrastive registration for unsupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Zhenbing Liu, Junfeng Hou, Xipeng Pan, Ruojie Zhang, and Zhenwei Shi. Pa-net: A phase attention network fusing venous and arterial phase features of ct images for liver tumor segmentation. *Computer Methods and Programs in Biomedicine*, 244:107997, 2024.

Shuichi Monzawa, Tomoaki Ichikawa, Hiroto Nakajima, Yuki Kitanaka, Kosaku Omata, and Tsutomu Araki. Dynamic ct for detecting small hepatocellular carcinoma: usefulness of delayed phase imaging. *American Journal of Roentgenology*, 188(1):147–153, 2007.

Yangfan Ni, Geng Chen, Zhan Feng, Heng Cui, Dimitris Metaxas, Shaoting Zhang, and Wentao Zhu. Da-tran: Multiphase liver tumor segmentation with a domain-adaptive transformer network. *Pattern Recognition*, 149:110233, 2024.

Farid Ouhmich, Vincent Agnus, Vincent Noblet, Fabrice Heitz, and Patrick Pessaux. Liver tissue segmentation in multiphase ct scans using cascaded convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14:1275–1284, 2019.

Shaohua Qiao, Mengfan Xue, Yan Zuo, Jiannan Zheng, Haodong Jiang, Xiangai Zeng, and Dongliang Peng. Four-phase ct lesion recognition based on multi-phase information fusion framework and spatiotemporal prediction module. *BioMedical Engineering OnLine*, 23(1):103, 2024.

Ashwin Raju, Chi-Tung Cheng, Yuankai Huo, Jinzheng Cai, Junzhou Huang, Jing Xiao, Le Lu, ChienHung Liao, and Adam P Harrison. Co-heterogeneous and adaptive segmentation from multi-source and multi-phase ct imaging data: A study on pathological liver and lesion segmentation. In *European Conference on Computer Vision*, pp. 448–465. Springer, 2020.

Sucheng Ren and Xiaomeng Li. Hresformer: Hybrid residual transformer for volumetric medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

J Gabriel Schneider, Zhen J Wang, Wilbur Wang, Judy Yee, Yanjun Fu, and Benjamin M Yeh. Patient-tailored scan delay for multiphase liver ct: improved scan quality and lesion conspicuity with a novel timing bolus method. *American Journal of Roentgenology*, 202(2):318–323, 2014.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

Youyi Song, Jeremy Yuen-Chun Teoh, Kup-Sze Choi, and Jing Qin. Dynamic loss weighting for multiorgan segmentation in medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10651–10662, 2023.

Changjian Sun, Shuxu Guo, Huimao Zhang, Jing Li, Meimei Chen, Shuzhi Ma, Lanyi Jin, Xiaoming Liu, Xueyan Li, and Xiaohua Qian. Automatic segmentation of liver tumors from multiphase contrast-enhanced ct images based on fcns. *Artificial Intelligence in Medicine*, 83:58–66, 2017.

Nikhil Kumar Tomar, Debesh Jha, Michael A Riegler, Håvard D Johansen, Dag Johansen, Jens Rittscher, Pål Halvorsen, and Sharib Ali. Fanet: A feedback attention network for improved biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9375–9388, 2022.

Bruce A Urban, Patricia A McGhie, and Elliot K Fishman. Helical ct: diagnostic pitfalls of arterial phase imaging of the upper abdomen. *American Journal of Roentgenology*, 174(2):455–461, 2000.

Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, 41(4):965–976, 2021.

Jinting Wang, Yujiao Tang, Yang Xiao, Joey Tianyi Zhou, Zhiwen Fang, and Feng Yang. Grenet: gradually recurrent network with curriculum learning for 2-d medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):10018–10032, 2023.

Yichao Wu, Qiang Zhou, Haoji Hu, Guanghua Rong, Yongwu Li, and Shiyan Wang. Hepatic lesion segmentation by combining plain and contrast-enhanced ct images with modality weighted u-net. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 255–259. IEEE, 2019.

Yingying Xu, Ming Cai, Lanfen Lin, Yue Zhang, Hongjie Hu, Zhiyi Peng, Qiaowei Zhang, Qingqing Chen, Xiongwei Mao, Yutaro Iwamoto, et al. Pa-resseg: A phase attention residual network for liver tumor segmentation from multiphase ct images. *Medical Physics*, 48(7):3752–3766, 2021.

Fan Zhang, Huiying Liu, Qing Cai, Chun-Mei Feng, Binglu Wang, Shanshan Wang, Junyu Dong, and David Zhang. Federated cross-incremental self-supervised learning for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 589–599. Springer, 2021a.

Yue Zhang, Chengtao Peng, Liying Peng, Huimin Huang, Ruofeng Tong, Lanfen Lin, Jingsong Li, Yen-Wei Chen, Qingqing Chen, Hongjie Hu, et al. Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 68–77. Springer, 2021b.

Yue Zhang, Chengtao Peng, Ruofeng Tong, Lanfen Lin, Yen-Wei Chen, Qingqing Chen, Hongjie Hu, and S Kevin Zhou. Multi-modal tumor segmentation with deformable aggregation and uncertain region inpainting. *IEEE Transactions on Medical Imaging*, 42(10):3091–3103, 2023.

Fuchen Zheng, Xinyi Chen, Xuhang Chen, Haolun Li, Xiaojiao Guo, Guoheng Huang, Chi-Man Pun, and Shoujun Zhou. Assnet: Adaptive semantic segmentation network for microtumors and multi-organ segmentation. *arXiv preprint arXiv:2409.07779*, 2024a.

Fuchen Zheng, Xinyi Chen, Xuhang Chen, Haolun Li, Xiaojiao Guo, Weihuang Liu, Chi-Man Pun, and Shoujun Zhou. Affsegnet: Adaptive feature fusion segmentation network for microtumors and multi-organ segmentation. *arXiv preprint arXiv:2409.07779*, 2024b.

W Zhong, F Liang, R Yang, and X Zhen. Prediction of microvascular invasion in hepatocellular carcinoma based on multi-phase dynamic enhanced ct radiomics feature and multi-classifier hierarchical fusion model. *Nan Fang yi ke da xue xue bao= Journal of Southern Medical University*, 44(2):260–269, 2024.

Yuyin Zhou, Yingwei Li, Zhishuai Zhang, Yan Wang, Angtian Wang, Elliot K Fishman, Alan L Yuille, and Seyoun Park. Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 155–163. Springer, 2019.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, pp. 3–11. Springer, 2018.

Xiner Zhu, Yichao Wu, Haoji Hu, Xianwei Zhuang, Jincao Yao, Di Ou, Wei Li, Mei Song, Na Feng, and Dong Xu. Medical lesion segmentation by combining multimodal images with modality weighted unet. *Medical Physics*, 49(6):3692–3704, 2022.

A ALGORITHM.

The overall workflow of the **MADF-Net** is summarized in algorithm 1.

Algorithm 1: High-level pseudocode of the proposed **MADF-Net**.

Input: Phase images: I_A, I_B, I_C; Clinical prior structure as a list of relations R; Reference mask T (only for training)

Output: Segmentation output Y_{out}

Initialize network parameters P; set mode flag isTrain
foreach *batch in training/eval* **do**
  /* basic preprocessing and branch init */
  $I'_A \leftarrow$ `PrepData(`I_A`)`
  $I'_B \leftarrow$ `PrepData(`I_B`)`
  $I'_C \leftarrow$ `PrepData(`I_C`)`
  $I_m \leftarrow$ CombineInit(I'_A, I'_B, I'_C) // early merged input for main branch
  BranchInputs $\leftarrow$ [I'_A, I'_B, I'_C, I_m]
  /* encode each branch to produce multi-level maps */
  foreach *entry J in BranchInputs* **do**
    FeatureMaps[J] $\leftarrow$ `EncodeBlock(`J`)` // returns list of maps at depths 1..D
  end
  /* clinical-aware message propagation (hierarchical depths) */
  for *depth d = 1* **to** *D* **do**
    for *node n in RelationOrder(*R*)* **do**
      Parents $\leftarrow$ GetParents(n, R)
      MsgIn $\leftarrow$ `PassMessages` ([FeatureMaps[parent][d] for parent in Parents])
      UpdatedMap[n][d] $\leftarrow$ UpdateUnit(FeatureMaps[n][d], MsgIn)
       // residual-style update
    end
  end
  /* per-phase local refinement and prepare temporal stack */
  for *phase p in* {*A,B,C*} **do**
    LowFeat $\leftarrow$ ExtractLow(UpdatedMap[p])
    Refined[p] $\leftarrow$ `LocalRefine(`*LowFeat*`)`
    StageOut[p] $\leftarrow$ ProjectForTemporal(Refined[p])
  end
  TemporalStack $\leftarrow$ Stack(StageOut[A], StageOut[B], StageOut[C])
  /* per-pixel temporal attention */
  TemporalEnhanced $\leftarrow$ `TemporalPerPixel(`*TemporalStack*`)`
  /* neighbor-aware cross-temporal fusion per phase */
  for *phase p in* {*A,B,C*} **do**
    Attended[p] $\leftarrow$ `NeighborInteract` (UpdatedMap[p], TemporalEnhanced)
     // neighbor queries + relative-pos bias
    Mix[p] $\leftarrow$ BlendLinear(Attended[p], $UpdatedMap[p]$) // weighted linear mixing
    FinalFeat[p] $\leftarrow$ `ChannelBoost` (Mix[p]) // channel gating / enhancement
end
/* merge multi-phase features and decode */
MergedFeat $\leftarrow$ `FinalMerge` (FinalFeat[A], FinalFeat[B], FinalFeat[C], UpdatedMap[I_m]) $Y_pred \leftarrow$ `DecodeBlock` (MergedFeat) // decoder with gated skips
/* auxiliary outputs aggregation (if enabled) */
AuxList $\leftarrow$ GetAuxOutputs() // possibly per-phase decoder heads
$Y_{\text{out}} \leftarrow$ OutputMerge($Y_pred, AuxList$)
/* loss and update (training only) */
if *isTrain* **then**
  LossVal $\leftarrow$ `CalcLoss` (Y_{out}, T)
  Backpropagate(LossVal, P)
end
return Y_{out}

Given the three single-phase CT images, the network first performs input-level preprocessing and branch initialization, generating both phase-specific and early-fused representations. These features are then processed through hierarchical encoder blocks, where clinical-prior message propagation is applied to capture inter-phase dependencies. Next, local refinement modules enhance low-level cues, followed by per-pixel temporal attention to model cross-phase temporal correlations. Neighbor-aware cross-temporal fusion and channel enhancement further integrate complementary information before multi-phase features are merged and decoded. Finally, decision-level aggregation combines auxiliary and main-branch outputs to produce the final segmentation mask. During training, the predicted mask is supervised by the ground-truth labels via a composite loss function.

B THE MPLL DATASET.

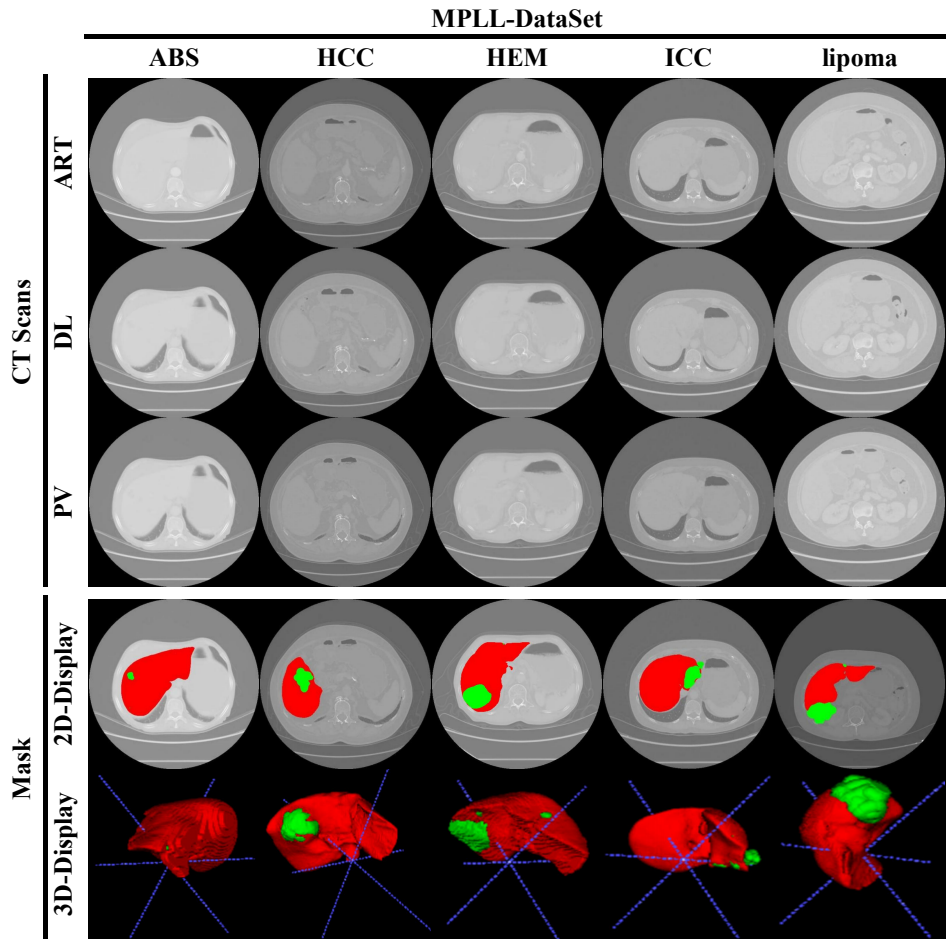

Figure 6: Example images from the MPLL dataset (red indicates liver regions, green indicates tumor regions).

❶: Patient Cohort and Imaging Protocol. The Multi-Phase Liver Lesion (MPLL) dataset was collected at the "Anonymous Authoritative Hospitals (information will be made public after the paper is accepted)". The dataset comprises $\mathbf{952,601}$ 2D slices, making it one of the largest publicly reported multi-phase CT resources for liver tumor segmentation research, comprising 141 patients diagnosed with a wide spectrum of hepatic diseases. Imaging was performed between 2018 and 2022, covering both pediatric and adult populations (ages 9–72 years). All cases underwent standardized multi-phase contrast-enhanced CT examinations that included arterial, portal venous, and delayed phases, thereby capturing complementary hemodynamic information. Each scan was acquired at an in-plane resolution of 512×512 pixels, while slice thickness ranged from 0.62 mm to 5.0 mm. Due to differences in anatomical coverage, the number of slices varied considerably across

patients (48–777 slices per study). Data were de-identified before release, with ethical approval obtained in advance.

❷: **Clinical Diversity and Pathology Spectrum.** MPLL was intentionally designed to reflect real-world clinical heterogeneity. It contains patients diagnosed with common malignant tumors such as hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma, alongside a range of benign lesions including cysts, hemangiomas, and abscesses. This diverse pathology coverage ensures that the dataset does not disproportionately represent a single disease entity, but instead provides a representative benchmark for developing algorithms that are robust across varying lesion types, morphologies, and enhancement characteristics.

❸: **Dataset Organization and Splitting Strategy.** To support reproducible research, the dataset was partitioned into training, validation, and testing cohorts following a 7:1:2 split protocol consistent with contemporary studies (Jiang et al., 2023). Importantly, the test set was fixed to 30 cases and completely withheld during model design and training, thereby guaranteeing unbiased evaluation. This design facilitates fair performance comparison across different methods and helps prevent information leakage during algorithm development.

❹: **Preprocessing and Annotation Pipeline.** One critical challenge of multi-phase imaging is the misalignment across arterial, portal venous, and delayed acquisitions caused by respiration, patient movement, or cardiac activity. To mitigate this, a B-spline deformable registration strategy was applied using the portal venous phase as reference. This procedure significantly reduces inter-phase variability and enables spatially consistent feature fusion. Ground-truth lesion masks were annotated in ITK-SNAP by two board-certified radiologists, followed by an adjudication step by a senior radiologist. This three-stage process was designed to maximize accuracy, reduce annotation bias, and enhance inter-observer agreement.

❺: **Comparative Advantages over Existing Datasets.** Unlike widely used liver CT datasets such as LiTS2017 (Bilic et al., 2023) and Medical Segmentation Decathlon (Task 3: Liver), which primarily focus on single-phase CT, MPLL offers multi-phase contrast-enhanced imaging across arterial, portal venous, and delayed phases. This temporal richness provides unique opportunities for investigating cross-phase fusion strategies, which are critical for accurate lesion delineation but are underexplored in existing benchmarks. Moreover, MPLL is substantially larger in terms of slice count (over 950k slices), contains a broader age range including pediatric cases, and offers a more diverse pathology spectrum that includes both malignant and benign liver lesions. The dataset therefore not only complements but also surpasses existing resources in its ability to support the development of clinically relevant and generalizable liver lesion segmentation methods.

❻: **Dataset Significance.** In summary, MPLL represents a large-scale, carefully curated, and clinically diverse benchmark for multi-phase liver lesion segmentation. Its strengths lie in the combination of temporal imaging information, broad pathology spectrum, rigorous preprocessing, and high-quality expert annotations. Together, these characteristics make MPLL an invaluable resource for advancing multi-phase fusion strategies in medical image analysis. Representative examples highlighting inter-phase contrast variations and lesion depiction are illustrated in Figure 2.

Table 5: Efficiency Comparison of `MADF-Net` and Baseline Models (GFLOPs and Parameters). Performance is evaluated on the MPLL dataset under the 3-phase experiment setting. The bold indicates the best.

Model	Gflops ($\times 10^9$)	Parameters (M)	Performance (%)
MAML (Zhang et al., 2021a)	**23.802**	4.216	75.49
MW-UNet (Zhu et al., 2022)	53.419	**2.773**	76.21
SA-Net (Zhang et al., 2021b)	152.965	170.852	76.29
PA-ResSeg (Xu et al., 2021)	64.660	67.732	77.17
MCDA-Net (Kuang et al., 2024)	89.480	48.717	77.40
Ours	99.552	40.482	**78.65**

C Detailed Efficiency Analysis.

As shown in Table 5, **MADF-Net** demonstrates a favorable balance between computational complexity, model capacity, and segmentation accuracy. While lightweight models such as MW-UNet achieve a small parameter size (2.773 M), they still require a non-trivial computational cost (53.419 GFLOPs) and their performance (76.21%) remains noticeably lower than ours. Similarly, MAML achieves the lowest GFLOPs (23.802) but suffers from a relatively limited accuracy (75.49%), which constrains its clinical applicability. On the other hand, heavier architectures like SA-Net and PA-ResSeg demand extremely large computational budgets (up to 152.965 GFLOPs and 170.852 M parameters), yet the corresponding accuracy (76.29% and 77.17%, respectively) provides only marginal improvement over lightweight baselines.

MADF-Net maintains a moderate parameter count of 40.482 M and a competitive computational demand of 99.552 GFLOPs, while delivering the **highest segmentation accuracy (**78.65%**)** among all compared methods. This clearly illustrates the *efficiency–accuracy trade-off*: although **MADF-Net** is not the most lightweight in terms of FLOPs or parameters, it achieves the best performance, outperforming both lightweight and heavyweight counterparts. This balance highlights **MADF-Net** 's practicality for real-world clinical deployment, where both computational feasibility and reliable accuracy are crucial.

D Large Language Models Usage Statement

LLMs were used only for language polishing in this work. The manuscript was drafted entirely by the authors, and LLMs were employed solely to refine grammar and clarity of English expression. All scientific ideas, methods, and results are original contributions of the human authors, with LLM assistance limited to post-writing editing akin to traditional proofreading.