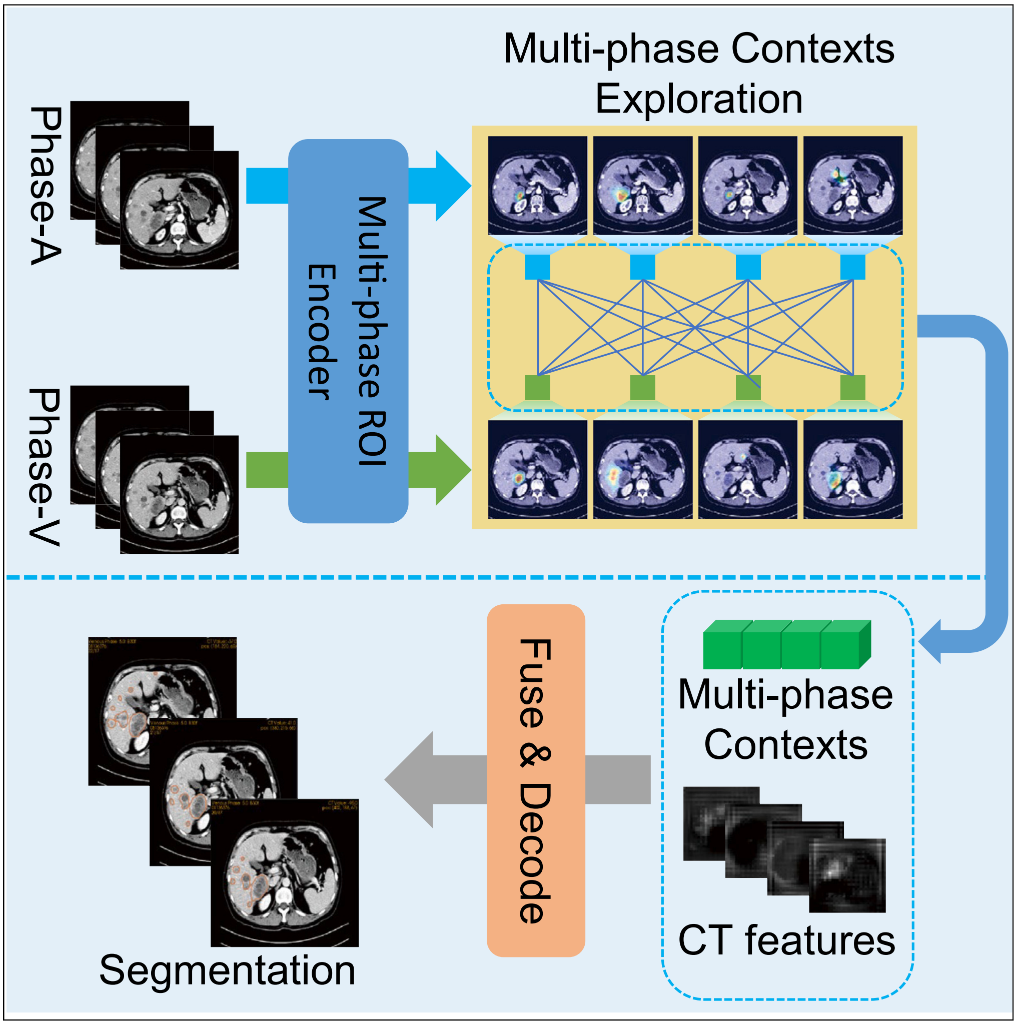
iScience

Article

Beyond radiologist-level liver lesion detection on multi-phase contrast-enhanced CT images by deep learning



Lei Wu, Haishuai Wang, Yining Chen, ..., Yuan Ding, Weilin Wang, Jiajun Bu

haishuai.wang@zju.edu.cn (H.W.)
dingyuan@zju.edu.cn (Y.D.)
wam@zju.edu.cn (W.W.)
bjj@zju.edu.cn (J.B.)

Highlights

MULLET achieved accurate segmentation of multi-phase CECTs for liver lesions

MULLET explores and utilizes multi-phase contexts in misaligned regions

MULLET performed significantly better than expert radiologists

Wu et al., iScience *26*, 108183
November 17, 2023 © 2023
https://doi.org/10.1016/j.isci.2023.108183

iScience

CellPress
OPEN ACCESS

Article

Beyond radiologist-level liver lesion detection on multi-phase contrast-enhanced CT images by deep learning

Lei Wu,[1,3,8] Haishuai Wang,[1,8,9,*] Yining Chen,[2,8] Xiang Zhang,[4,5] Tianyun Zhang,[6] Ning Shen,[6] Guangyu Tao,[7] Zhongquan Sun,[2] Yuan Ding,[2,*] Weilin Wang,[2,*] and Jiajun Bu[1,*]

SUMMARY

Accurate detection of liver lesions from multi-phase contrast-enhanced CT (CECT) scans is a fundamental step for precise liver diagnosis and treatment. However, the analysis of multi-phase contexts is heavily challenged by the misalignment caused by respiration coupled with the movement of organs. Here, we proposed an AI system for multi-phase liver lesion segmentation (named MULLET) for precise and fully automatic segmentation of real-patient CECT images. MULLET enables effectively embedding the important ROIs of CECT images and exploring multi-phase contexts by introducing a transformer-based attention mechanism. Evaluated on 1,229 CECT scans from 1,197 patients, MULLET demonstrated significant performance gains in terms of Dice, Recall, and F2 score, which are 5.80%, 6.57%, and 5.87% higher than state of the arts, respectively. MULLET has been successfully deployed in real-world settings. The deployed AI web server provides a powerful system to boost clinical workflows of liver lesion diagnosis and could be straightforwardly extended to general CECT analyses.

INTRODUCTION

The precision diagnosis and prediction of liver cancer are fundamental to clinical practice in diagnosing and treating liver tumors.[1–3] For liver lesion screening, contrast-enhanced computed tomography (CECT) has been clinically proven effective to assist diagnosis and surgery.[4–6] Yet, it is time-consuming when manually performed on numerous CECTs.[7–9] With the advent of the artificial intelligence (AI) era, cutting-edge computer technology (e.g., semantic segmentation[9–11]) has shown an exciting potential to segment and identify liver lesions from CECT images.[12–18] Hence, segmenting and classifying liver lesions from CECT images by leveraging an efficient AI model is crucial to assist doctors in liver lesion diagnosis.

Multi-phase textures (i.e., the hepatic arterial phase (phase-A) and the venous phase (phase-V) of CECT) are essential to detect and diagnose liver lesions.[4–6,19,20] The abundant multi-phase contextual information is also conducive to training AI models for liver lesion segmentation.[12,14,21] However, there are still two main challenges to exploring the multi-phase contexts in liver lesion segmentation tasks. First, the respiration and the movement of abdominal organs cause considerable liver movements and deformation during multiple scans, resulting in pixel-wise misalignment between different phases. Thus, the misalignment increases the challenge of leveraging multi-phase contexts for misaligned regions,[22,23] especially for small lesions. Second, previous efforts suffer from poor performance as well as limited generalizability because they are typically trained on small datasets (i.e., 30–400 scans).[13,14,24–26] However, fully automatic liver lesion segmentation usually requires a large quantity of CECT images (e.g., more than 1,000 scans) for training in an end-to-end fashion.[27]

In the past decade, several pioneer methodologies, e.g., intensity threshold[28] and region growing,[22,29] were explored for medical image segmentation. Nevertheless, these methods highly rely on handcrafted or low-level features, which are sensitive to complicated abdominal CECT images and have limited feature representation capability.[12,16,17,30] The emerging deep neural networks have recently demonstrated promising performance in the precise diagnosis and treatment of various diseases automatically via CT images,[15–17,21,27,31,32] thanks to the strong ability to learn high-level feature representations.[7,17,18,33,34] For multi-phase CECT images, a majority of the existing deep-learning-based methods either ignore multi-phase contexts[15,16,27,35] or linearly transform (e.g., simply concatenation or pixel-wise feature filtering)

[1]Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou, China
[2]Department of Hepatobiliary and Pancreatic Surgery, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3]Pujian Technology, Hangzhou, Zhejiang, China
[4]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA
[5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[6]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, Zhejiang, China
[7]Department of Radiology, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[8]These authors contributed equally
[9]Lead contact
*Correspondence: haishuai.wang@zju.edu.cn (H.W.), dingyuan@zju.edu.cn (Y.D.), wam@zju.edu.cn (W.W.), bjj@zju.edu.cn (J.B.)

https://doi.org/10.1016/j.isci.2023.108183

Table 1. Description and characteristics of the AIELDI dataset

Demographics	
# of CECT	1,229
# of patient	1,197
Female	454 (37.93%)
Male	743 (62.07%)
Age	58.96 (3–90)
Image protocols	
Resolution (mm)	0.472–0.890
Manufacturer	SIEMENS (1,184); GE MEDICAL SYSTEMS (24); TOSHIBA (21)
Manufacturer's model	SOMATOM Definition AS; SOMATOM Definition AS+; SOMATOM Definition Flash; Optima CT540; Aquilion ONE; Sensation 16; SOMATOM Force
Thickness	3.0–5.0
Slice range	32–116

There are 1,229 multi-phase CECTs that contain 4,660 lesions and 155,766 CT slices. Each CECT scan contains an artery phase and a venous phase, and each phase consists of a sequence of CT slices. The age variable is presented as average (range). The resolution variable is presented as a range, and the manufacturer is presented as name (count).

multi-phase CT images to explore multi-phase contextual information.[25,26,31,36] Assuming liver lesions are well aligned between different phases, the linear transform strategies can effectively capture the cross-phase contexts of a feature point in one phase from the corresponding feature point in the other phase. However, they fail to explore and utilize the multi-phase contexts, which are crucial in misaligned liver lesion regions, especially for small lesions. Hereby, we developed an end-to-end AI system named multi-phase liver lesion segmentation (MULLET) to automatically and accurately segment liver lesions directly from multi-phase CECT images. Specifically, MULLET first utilized a specific pre-trained convolutional neural network (CNN) to extract features of each slice in the multi-phase CECT images. Afterward, a multi-phase region of interest (ROI) embedding module leveraged both global and local features to produce a range of ROIs for each CT slice. Each generated ROI was then embedded to a fixed-dimensional embedding. Our multi-phase ROI encoder aims to emphasize prominent regions of the CECT images and reduce the computational complexity of various operations. To this end, we employed two multi-head attention[37] modules to aggregate multi-phase contexts by computing the latent relationship between ROI embeddings of different phases. Finally, MULLET reverted multi-phase contexts to the original size of ROI features and produced final segmentation by a decoder.[34]

MULLET achieved promising performance in segmenting liver lesions from multi-phase CECT images. To validate that our AI system is clinically stable and precise, we evaluated it on a private dataset that contains 1,229 multi-phase CECT scans, to the best of our knowledge, which is the largest multi-phase CECT dataset for liver lesion segmentation. To further demonstrate the generalization capability of our AI system, we applied it to two external benchmarking datasets (i.e., BRATS and CHAOS) and also achieved superior performance compared with state-of-the-art studies. Importantly, we deployed the proposed system at The Second Affiliated Hospital Zhejiang University School of Medicine in March 2021. Our system can analyze 79,949 CT images annually and is expected to save 13,324 h of healthcare professionals in lesion scan segmentation each year. Our MULLET suggests an excellent potential for advanced clinically applicable AI systems to tackle the limitations and segment liver lesions from CECT images.

RESULTS

The AIELDI dataset

We launched an AI-Empowered Liver Diagnostic Initiative (AIELDI) that collected a total of 1,229 multi-phase CECT scans (1,197 patients; 454 females and 743 males; age 58.96 ± 11.7; Table 1) at The Second Affiliated Hospital Zhejiang University School of Medicine. To our knowledge, this is the largest multi-phase CECT dataset for liver lesion segmentation. All the multi-phase CECT scans were exported from the picture archiving and communication system[38] of the hospital and stored in a Digital Imaging and Communications in Medicine (DICOM) format.[39] For annotation, two experts with five years of experience labeled the biopsied or CECT images first, and then two experts with 10+ years of experience reexamined the results. The expert with 23 years of experience will make the final annotation decision (a.k.a., gold standard or ground truth) if there are discrepancies between the two experts with 10+ years of experience. The annotated liver CECT images were randomly divided into a training set of 1,029 multi-phase CECT images and a test set of the remaining 200 samples to develop our AI system. This dataset consists of 1,983 malignant tumors, 418 hemangiomas, and 2,259 cysts by the category range, 3,133 small lesions (≤ 1 cm), 537 medium lesions (≤ 1.5 cm), and 990 large lesions (> 1.5 cm) by maximum diameter range, as shown in Table S1.

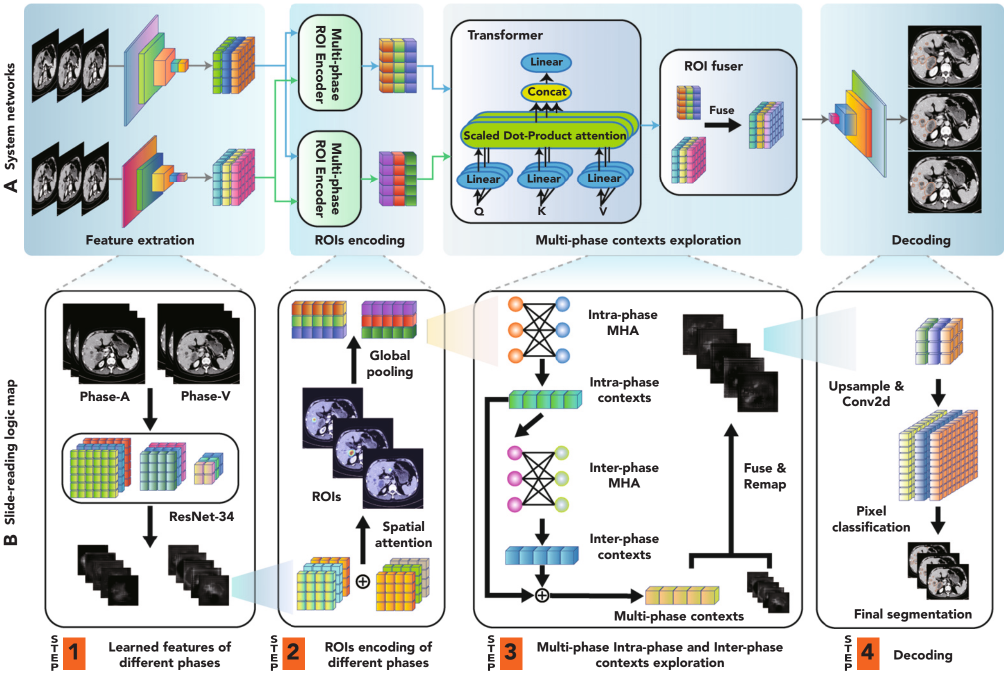


Figure 1. The pipeline of our proposed AI system for segmenting liver lesions from CECT images

(A) The system consists of four major steps, including feature extraction, ROIs encoding, multi-phase contexts exploration, and decoding.

(B) The detailed architecture of each component. In the first step, a CNN-based network (e.g., ResNet-34) was used to extract features from a multi-phase CECT. In the second step, a multi-phase region of interest (ROI) encoder extracted both global and local features via attention mechanisms, followed by a spatial attention module to produce several ROIs for each CECT slice. The information from all ROIs was encoded in a set of low-dimensional feature representations (vectors). In the third step, we adopted a transformer-based multi-phase context exploration module to explore the inter-phase and intra-phase relationship between the ROIs of different phases. Finally, a simple decoder was applied to produce the liver lesion segmentation.

MULLET enabled effective multi-phase information exploration and integration in misaligned regions

The developed model aims to effectively and accurately segment liver lesions using multi-phase CECT scans. To this end, the premise of improving performance by MULLET follows two propositions: (i) the model learned comprehensive features by incorporating cross-phase global and local information from multi-phase CECT images; and (ii) the model alleviated the effect of misalignment between phases by adopting a better exploration of multi-phase contexts at misaligned regions of liver lesions.

Given the intrinsic nature of multi-phase abdominal CECT, the model workflow consists of four steps (Figure 1). First, we extracted distinct features of all slices in all phases through a CNN-based feature extraction module.[40,41] Second, we proposed a multi-phase ROI encoder (Figure 2) that extracts cross-phase global and local features (Figure 3) via attention mechanisms. The cross-phase features were then utilized to guide a spatial attention module to produce several ROIs for each CECT slice.[42] Subsequently, we adopted a transformer[37] to explore the inter-phase and intra-phase relationship between the ROIs. In the final step, we employed a simple yet effective decoder[34] to segment the lesion images.

MULLET achieved accurate segmentation of multi-phase CECTs for liver lesions

To assess the performance of liver lesion segmentation, multi-phase CECT scans with gold-standard expert annotation were chosen as the benchmarks. Taking a multi-phase CECT scan that contains phase-A and phase-V as the input, our framework utilized cross-phase features for liver lesion segmentation (see Model implementation, STAR Methods section). We adopted the commonly used metrics, i.e., Dice per case (DPC), volumetric overlap error (VOE), recall, precision, and F2-score, to comprehensively evaluate the segmentation performance (see Evaluation metrics in STAR Methods). The VOE metric is the lower, the better, while other metrics are the higher and the better. We further conducted experiments for the evaluation in terms of different lesion sizes and categories. Large lesions tend to yield more simple and convincing segmentation, while small lesions are typically difficult and imprecise. However, the segmentation of small lesions is more urgent in clinical diagnosis because diagnosing small lesions is more time-consuming for physicians. The ability to deal with small lesions is thus particularly important for liver lesion segmentation. Therefore, we systematically evaluated the proposed MULLET from three aspects: (1) the general performance with respect to the overall distribution of the dataset, (2) the performance with respect to various lesion sizes, especially small lesions, and (3) the performance with respect to various lesion categories.

First, we evaluated the proposed model over all CECTs in the AIELDI dataset regardless of lesion sizes or categories. Our AI system achieved a DPC score of 78.47%, a VOE value of 32.23%, a recall of 89.90%, a precision of 80.48%, and an F2 score of 87.83% in terms of

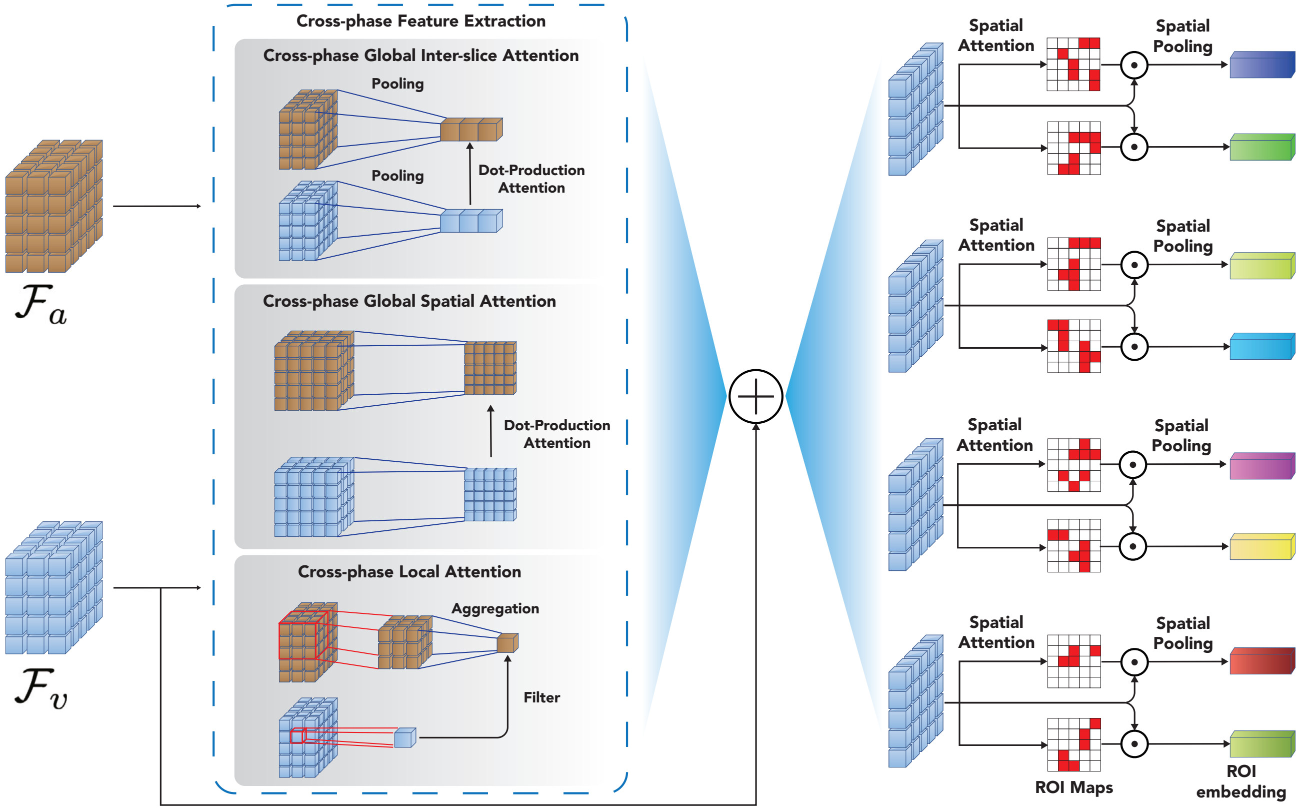


Figure 2. The architecture of the multi-phase ROI encoder

In the cross-phase global inter-slice attention module, we squeezed the features by spatial pooling and extracted the cross-phase features along the z axis. Similarly, the cross-phase global spatial attention module pooled the features along the axial plane and computed cross-phase spatial features. In the cross-phase local attention module, a Conv3D was used to aggregate the corresponding neighboring information for each pixel. A filtering map was then computed by an element-wise multiplication and a sigmoid function to filter the neighboring information.

segmenting liver lesions, as shown in Figure 4A. Our AI system significantly outperformed existing approaches and expert radiologists for liver lesion segmentation from complex misaligned multi-phase CECT images (Tables 2 and 3; Figure 5). Next, we evaluated the proposed model within each category of liver lesions. Figure 4B demonstrates the performance of MULLET in the categories of malignant, hemangioma, and cyst. It can be observed that MULLET achieved an F2 score of 90.88% and 90.63% in the malignant tumor and hemangioma, respectively. The F2 score in the cyst category is 81.45%, which is lower than in other categories. This is because most cysts are very small and misaligned between phases, resulting in difficulty in detecting cysts.

Further, liver lesions were divided into small, medium, and large groups according to the maximum diameter of lesions, i.e., small group ($\leq$ 1 cm), medium group ($\leq$ 1.5 cm), and large group (> 1.5 cm). As shown in Figure 4C, our proposed MULLET achieved excellent performance on large lesions, i.e., a recall of 98.53%, a precision of 96.92%, and an F2 score of 98.20%. For medium lesions, MULLET still achieved a recall of 91.62%, precision of 92.91%, and F2 score of 91.86%. The performance in the small group is obviously lower than in other groups (F2 score of 82.41%) because small lesions are more difficult to segment and tend to be ignored in clinical settings. In addition, small lesions are more sensitive to the misaligned regions. Despite the practical challenges, the performance of our AI system for small lesions still achieved acceptable performance and outperformed other existing competitive methods (Section Systematic benchmarking demonstrates superior performance compared to existing tools), suggesting the great ability of our model in dealing with various styles of liver lesions. This is extremely important for liver lesion segmentation in real-world clinical practice.

Interestingly, we observed that the performance varied in different gender and age groups, though the distribution of different groups is identical, see also in Table S2. From Figure 4D, we can see that our AI system performed better in the female group, with gaps of 2.85% and 4.61% in terms of DPC and F2 score. As shown in Figure 4E, our AI system achieved a better performance in the age group under 65, with gaps of 2.08% and 5.11% in terms of DPC and F2 score.

In addition to quantitative evaluation, Figure 4F also presents a qualitative evaluation, illustrating the representative segmentation produced by MULLET for small cysts, large malignant tumors, and large hemangioma. As shown in the first case, our MULLET was able to segment all small cysts accurately. MULLET also yielded accurate boundaries for a large malignant tumor in the second case, even though it hangs out of the liver. In the third case, MULLET segmented accurate boundaries for the large hemangioma. As expected, we found

Table 2. Quantitative comparison between our method and state of the arts on AIELDI

Methods	DPC (%)	VOE (%)	Recall (%)	Precision (%)	F2 Score
MC-FCN	71.60±0.51	40.00±0.48	78.84±1.11	83.93±1.18	79.81±0.33
MMNet	63.74±5.96	48.31±6.35	73.49±6.27	87.12±4.13	75.82±2.09
HRNet	69.07±0.95	42.56±0.90	79.35±1.21	64.37±4.02	75.80±0.51
M^3Net	70.62±0.35	40.81±0.34	77.94±1.53	80.72±0.87	78.48±0.55
nnFormer	70.08±0.81	40.76±0.78	78.52±2.28	84.59±3.52	79.65±1.42
nn-UNet	70.93±0.75	41.47±0.88	86.91±4.22	57.97±10.1	78.83±1.42
SAM	72.17±1.29	39.04±1.50	80.83±2.01	86.88±1.13	81.96±0.81
PA-ResSeg	72.67±0.73	38.80±0.39	83.87±0.65	71.89±1.96	81.16±0.92
MULLET	**78.47±0.51**	**32.23±0.88**	**89.90±1.40**	80.48±3.00	**87.83±1.20**

Our model significantly outperformed all the competing methods for the segmentation of liver lesion boundaries, i.e., an improvement of 5.80% (p value = 2.37e-8) and 6.57% (p value = 9.5e-10) in terms of DPC and VOE comparing to PA-ResSeg, respectively. In addition, our model achieved a comprehensive improvement for liver lesion classification, e.g., an improvement of 5.87% (p value = 1.9e-10) in terms of F2 score compared to SAM.

that, although the lesion styles and sizes vary highly across different categories and groups, our MULLET still accurately and robustly segmented various liver lesions (comprehensive qualitative results are detailed in Figures S2–S4).

To interpret MULLET's segmentation results, we highlighted regions of high importance according to the weight maps of ROIs in the multi-phase ROI encoder, which are visualized in Figure 6D. We can observe that most ROI weight maps highlighted suspicious lesion regions in the multi-phase CECT images, indicating that our AI system enabled the identification of discriminative features driving liver lesion segmentation. Remarkably, it is consistent with clinical diagnoses by doctors that only distinct concern regions correlate with specific disease categories, suggesting the trustworthy interpretability of our AI system.

Systematic benchmarking demonstrates superior performance compared to existing tools

To evaluate the superior performance compared with existing methods, we further benchmarked MULLET against multiple state-of-the-art deep-learning-based methods for liver lesion segmentation, including MC-FCN,[25] M^3 Net,[23] MMNet,[21] nn-UNet,[18] nnFormer,[43] HRNet,[44] SAM,[14] and PA-ResSeg[13] (see detailed results in the STAR Methods section). Note that MC-FCN incorporates multiphase contexts for liver lesion segmentation by stacking single-phase 2D images in the input layer or merging high-level features produced by 2D networks independently for each phase. M^3 Net adopts a 3D residual network that compresses the 3D input tensor into 2D features and a Cross-phase Non-local Attention module to relieve the misalignment between phases. Other approaches (i.e., MMNet, SAM, and PA-ResSeg) propose a powerful cross-phase fusion mechanism to fulfill the interaction between the features extracted from multiple phases. We also provided p values (two-sided t test) to validate the statistical significance of the improvements. The statistical significance is defined as 0.05.

As shown in Table 2, we have two important observations with respect to the comparison results on the AIELDI dataset: (i) our model significantly outperformed all the competing methods with respect to the segmentation of liver lesion boundaries, i.e., an improvement of 5.8% (p value = 2.37e-8) and 6.57% (p value = 9.5e-10) in terms of DPC and VOE comparing to PA-ResSeg (the best baseline), respectively; and (ii) our model achieved a comprehensive improvement for liver lesion classification, e.g., an improvement of 5.87% in terms of F2 score compared to SAM (p value = 1.9e-10). Although MF-FCN and MMNet performed better in terms of precision, the inferior performance on recall limits the value of their clinical application.

Additionally, it can be seen from Figure 5 that, compared with baseline methods, our AI system has the most improved comprehensive performance with respect to different groups divided by the lesion categories or lesion sizes. Specifically, compared to the best performance in competitive methods, the lesion segmentation performance (Figure 5A) was improved by 3.60%, 12.45%, and 2.22% on F2 score in the group of malignant tumor, hemangioma, and cyst, respectively. For the three categories, the paired p values of F2 score were 4.1e-5, 1.6e-6, and 2.4e-2. Although the segmentation performance for large lesions was not improved significantly, MULLET achieved remarkable enhancement of the performance for small lesions, with improvements of 1.52% (compared with nn-UNet) and 6.92% (compared with SAM) in terms of recall and F2 score as shown in Figure 5B. In addition, compared with PA-ResSeg, MULLET also achieved improvements of 4.22% for medium lesions in terms of F2 score. For small and medium lesions, the paired p values of F2 score were 9.6e-6 and 0.007. The results turned out to be the advancement of various strategies we proposed. For example, instead of information filters or inductive bias used in existing methods, MULLET learned several ROIs for each CT slice and adopted transformer to explore multi-phase contexts without any inductive biases. Consequently, MULLET enabled more effective handling of complex scenarios (e.g., severely misaligned small tumors). More experiments and results on publicly available datasets are reported in Figure S1 to investigate the effects of different misalignment levels for the AI models.

The comparisons of DCP in different groups (i.e., lesion category, lesion size, and age) are shown in Figure 5C. MULLET outperformed nnFormer and SAM in terms of the segmentation of malignant tumors and hemangiomas, with improvements of 6.16% and 1.80%, respectively. However, PA-ResSeg and SAM performed slightly better than MULLET. For the other two groups (i.e., lesion size and age), our MULLET consistently performed better than other baseline methods.

Table 3. Quantitative comparison between our AI system and two expert radiologists

	Metrics	Expert-1	Expert-2	AI
Malignant	DPC	53.13	74.64	**85.17**
	Recall	76.22	83.06	**92.18**
	Precision	95.51	91.07	82.99
	F2 Scores	79.43	84.55	**90.18**
Hemangioma	DPC	71.05	79.86	**82.31**
	Recall	61.90	64.29	**76.19**
	Precision	68.42	79.41	69.57
	F2 Scores	63.11	66.83	**74.77**
Cyst	DPC	73.07	73.69	**75.29**
	Recall	70.21	71.63	**90.78**
	Precision	80.49	80.80	64.97
	F2 Scores	72.05	73.29	**84.10**
Small	DPC	48.62	60.69	**63.01**
	Recall	67.41	73.42	**87.66**
	Precision	86.59	85.61	68.23
	F2 Scores	70.53	75.57	**82.93**
Medium	DPC	82.46	85.74	**87.24**
	Recall	81.67	88.33	**90.00**
	Precision	90.74	89.83	85.71
	F2 Scores	83.33	88.63	**89.11**
Large	DPC	81.66	82.39	**84.30**
	Recall	85.09	85.96	**98.23**
	Precision	91.51	89.91	**97.39**
	F2 Scores	86.30	86.73	**98.07**
Overall performance	DPC (L65)	73.20	74.09	**79.07**
	DPC (G65)	76.34	76.95	**79.78**
	DPC	73.79	74.70	**78.89**
	VOE	37.52	36.68	**32.12**
	Recall	73.27	78.16	**90.40**
	Precision	88.42	87.24	75.86
	F2 Score	75.87	79.82	**87.07**
	Time (min)	15.00	14.80	**0.34**
	Time (min) (with AI assistance)	4.8	5.0	–

We tested on 100 CECT scans randomly selected from the AIELDI test dataset for liver lesion segmentation. – means the value is not applicable.

As a qualitative evaluation, we show the representative segmentation produced by different methods on the testing CECT images in Figure 6A, where different colors mark the malignant tumors, hemangiomas, and cysts. Compared with other deep-learning-based models, we observed that our AI system yielded more accurate segmentation boundaries of liver lesions and was more robust on small tumors. Compared to the best baseline (PA-ResSeg), our AI system stably found the small lesions (i.e., small cysts and small malignant tumors in Figure 6A) and achieved perfect boundaries of large lesions (i.e., large malignant tumor and large hemangiomas in Figure 6A).

Comparison with human experts

In addition, to compare with advanced deep learning methods, we further compared the performance of our AI system with expert radiologists. Two radiologists with 5 and 10 years of clinical experience independently segmented 100 multi-phase CECT scans randomly selected from our test dataset of liver lesions. Note that they did not participate in ground-truth label annotation.

Table 3 shows our AI system performed significantly better than both expert radiologists, with improvements of 5.1% (radiologist 1) and 4.19% (radiologist 2) in terms of DPC, the improvements of 5.4% (radiologist 1) and 4.56% (radiologist 2) in terms of VOE, and the improvements of 11.20%

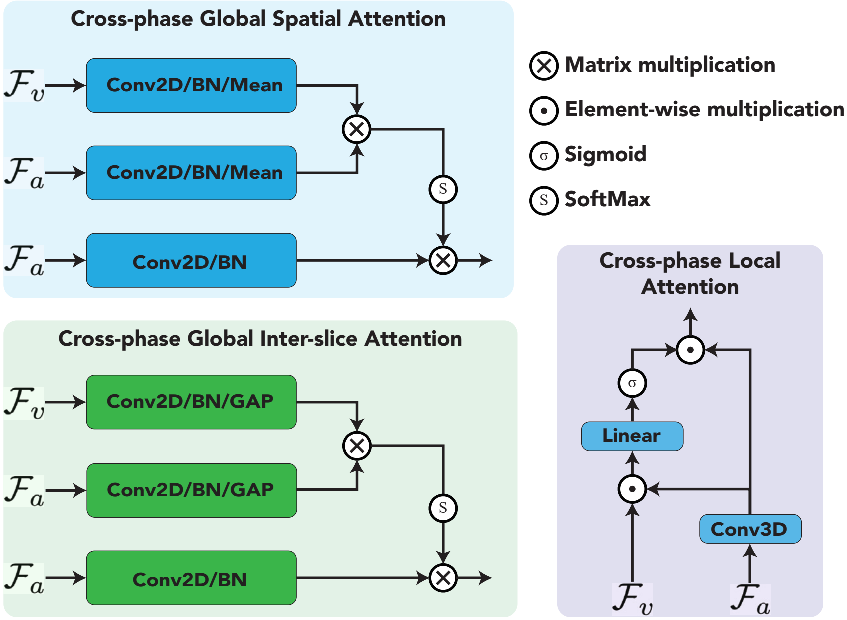


Figure 3. The architectures of cross global spatial attention (CGSA), cross global inter-slice attention (CGIA), and cross local attention (CLA)

In CGSA, we squeezed the information along z axis and computed Dot-Production attention between different phases. Similarly, CGIA squeezed the features by a spatial pooling and computed intra-slice relationship between different phases. In CLA, we applied Conv3D to gather neighboring information for each pixel and computed a gate map to filter cross-phase local information.

(radiologist 1) and 7.25% (radiologist 2) in terms of F2 score. Although the two radiologists performed better than our AI system on precision, their recall was significantly worse than ours. Specifically, in terms of recall, our AI system achieved the improvements of 20.25% (radiologist 1) and 14.24% (radiologist 2) for small lesions, the improvements of 8.33% (radiologist 1) and 1.67% (radiologist 2) for medium lesions, and the improvements of 13.14% (radiologist 1) and 12.27% (radiologist 2) for large lesions. For different types of lesions, our AI system also significantly outperformed two radiologists in terms of F2 score, with improvements of 10.75% (radiologist 1) and 5.63% (radiologist 2) for malignant tumors, the improvements of 11.66% (radiologist 1) and 7.94% (radiologist 2) for hemangiomas, and the improvements of 12.05% (radiologist 1) and 10.81% (radiologist 2) for cysts. Furthermore, it is worth noting that expert radiologists' segmentation efficiency was significantly worse than our AI system. For example, two radiologists spent 15.0 and 14.8 min (on average) annotating one CECT scan, as shown in Table 3. In contrast, our AI system could segment one CECT scan in only 20.59 s, which is 43.4 times faster than domain experts on average.

Since our AI system experimentally achieved superior performance in terms of efficiency and accuracy, we were also interested in investigating whether our system can facilitate clinical workflow and improve the efficiency of clinical diagnosis. Therefore, we applied the deployed AI system to assist radiologists with liver lesion segmentation rather than fully manual segmenting from scratch. Specifically, we adopted our system to generate initial segmentation, and radiologists were only required to check and amend the results produced by the AI system. As shown in Table 3, with assistance from our AI system, the annotation time was reduced from 14.9 min to about 4.9 min on average, a 67.1% reduction in segmentation time.

External evaluation on public datasets

We applied our AI system to two public datasets with 5-fold cross-validation for external testing to investigate our model's robustness and generalization ability. The two open datasets are BraTS2020[45] and CHAOS.[46] We only considered the task of MRI segmentation for abdominal organs in CHAOS. Since the phases in BraTs are well aligned, we introduced noise (i.e., scaling and shifting) on T1CE and T2 to simulate the misalignment. Table 4 shows that our AI system achieved the Dice scores of 73.44%, 80.70%, and 77.10% for enhanced tumor, tumor core, and whole tumor, respectively. Compared with the best baseline nn-UNet, the average DPC improvement is 1.29% (p values = 2.9e-4).

Table 5 shows our AI system significantly improved the segmentation of the right kidney, left kidney, and spleen, i.e., an improvement of 6.48%, 7.64%, and 7.98%, respectively. Specifically, for the other three categories, the paired p values are 2.3e-5, 6.8e-4, and 3.2e-4, respectively. Overall, all the results demonstrated the superior performance of our system on the public datasets, suggesting the high robustness and generalization ability of MULLET.

Webserver

Due to the privacy and security of the data, we deployed the off-the-shelf system offline at The Second Affiliated Hospital Zhejiang University School of Medicine for further retrospective studies on March 1st, 2021. We provide a guest account (account: 13611112222; password:

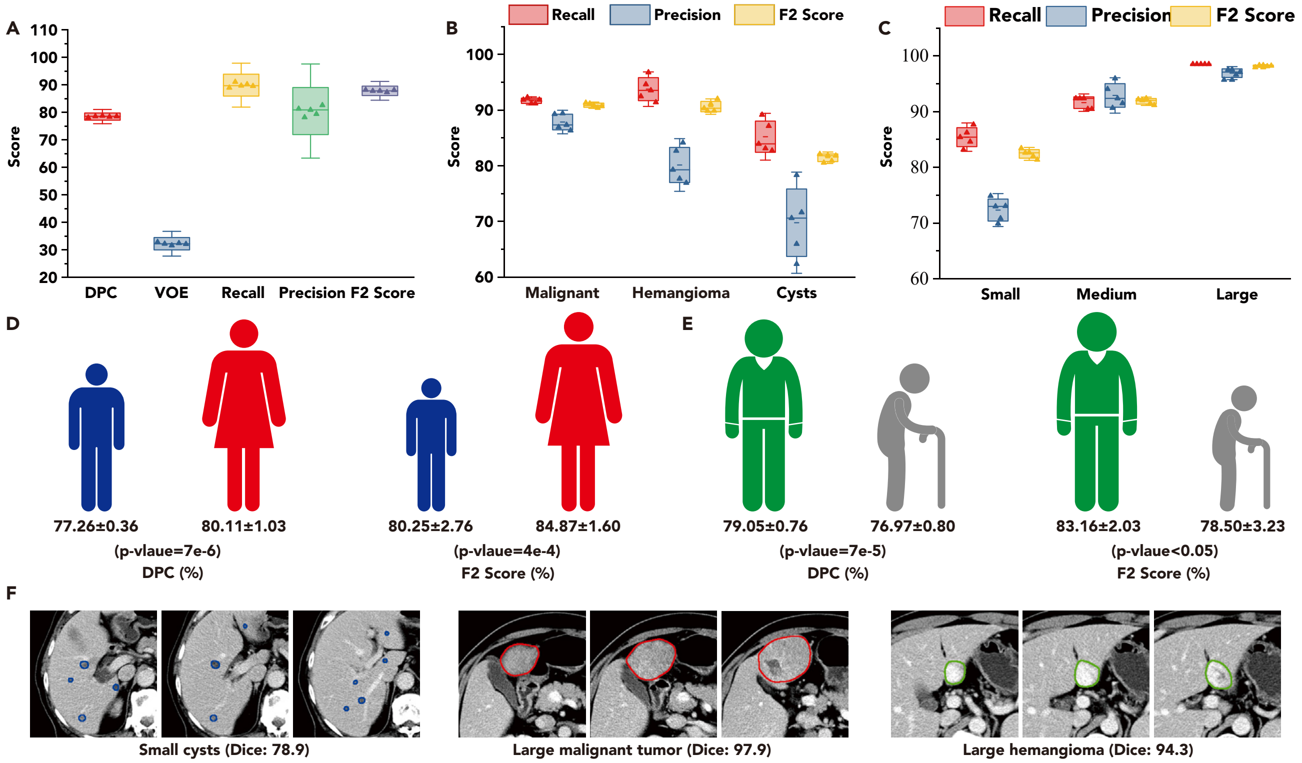


Figure 4. Quantitative and qualitative evaluations of our AI system with respect to various types of liver lesions

Data are represented as mean ± SEM. The statistical information of liver lesions is presented in Table S1. The ranges of box and whisker are standard deviations.

(A) Overall performance of our AI system for liver lesion segmentation. The coefficients of the box and whisker are 5 and 10 for better visualization. Our AI system achieved a DPC score of 78.47%, a VOE value of 32.23%, a recall of 89.90%, a precision of 80.48%, and an F2 score of 87.83% in segmenting liver lesions. Our AI system significantly outperformed existing approaches and expert radiologists for liver lesion segmentation from complex misaligned multi-phase CECT images (Tables 2 and 3; Figure 5).

(B) The performance with respect to various categories of liver lesions. The coefficients of the box and whisker are 1 and 1.5 for better visualization. It can be observed that MULLET achieved F2 scores of 90.88% and 90.63% for segmenting the malignant tumor and hemangioma, respectively. The F2 score in the cyst category is 81.45%, which is lower than in other categories. This is because most cysts are typically very small and highly misaligned between phases, causing difficulty in detecting cysts.

(C) The performance with respect to various sizes (i.e., maximum diameter) of liver lesions. The coefficients of the box and whisker are 2 and 4 for better display, respectively. Our proposed MULLET achieved excellent performance on large lesions, i.e., recall of 98.53%, precision of 96.92%, and F2 score of 98.20%. For medium lesions, MULLET still achieved a recall of 91.62%, precision of 92.91%, and F2 score of 91.86%. The performance in the small lesion group decreased dramatically (i.e., F2 score of 82.41%) because small lesions are more difficult to segment and tend to be ignored in clinical settings.

(D) Performance by gender. Our AI system performed better in women than in men, with gaps of 2.85% and 4.61% in terms of DPC and F2 scores. The inter-gender disparity is not biased by the data distribution, evidenced by the predominant role of males in the dataset (64.07% patients are male; Table 1).

(E) Performance by age. Our AI system achieved higher performance among adults (<65) than elders (≥ 65) with gaps of 2.08% and 5.11% in terms of DPC and F2 score.

(F) The segmentation produced by our AI system. We visualized the adjacent three CT slices in phase-V of three cases. Our MULLET could locate small cysts and segment them accurately in the first case. As shown in the second case, MULLET could circle boundaries accurately for the large malignant tumor, even though it is out of the liver. In the third case, MULLET also segmented accurate boundaries for large hemangioma. See also Figure S1.

Pj123456) to the online system (http://liver.uiiai.com). Fourteen de-identified patients (8 females and 6 males) are provided as examples. As of June 30th, 2022, our MULLET system has analyzed 126,587 CT images, most containing no focal liver lesions. According to the observer study, our system can analyze 79,949 CT images annually and is expected to save 13,324 h of healthcare professionals in lesion scan segmentation each year. The system requires users to submit a zip file that all files are in .dicom format to be fed into the proposed MULLET model. The AI system will automatically launch the inference task after receiving the input file. Users can check the analytic results yielded from the trained model through the web interface.

DISCUSSION

AI-empowered liver lesion segmentation is essential in helping develop an automatic and accurate approach for clinical diagnosis in digital medicine. It is still a fundamental challenge to exploit multi-phase CECT images for liver lesion diagnosis, where the misalignment between phases and the fuzzy edge of small lesions is still not well explored and addressed. In this work, we introduced a fully automatic and clinically

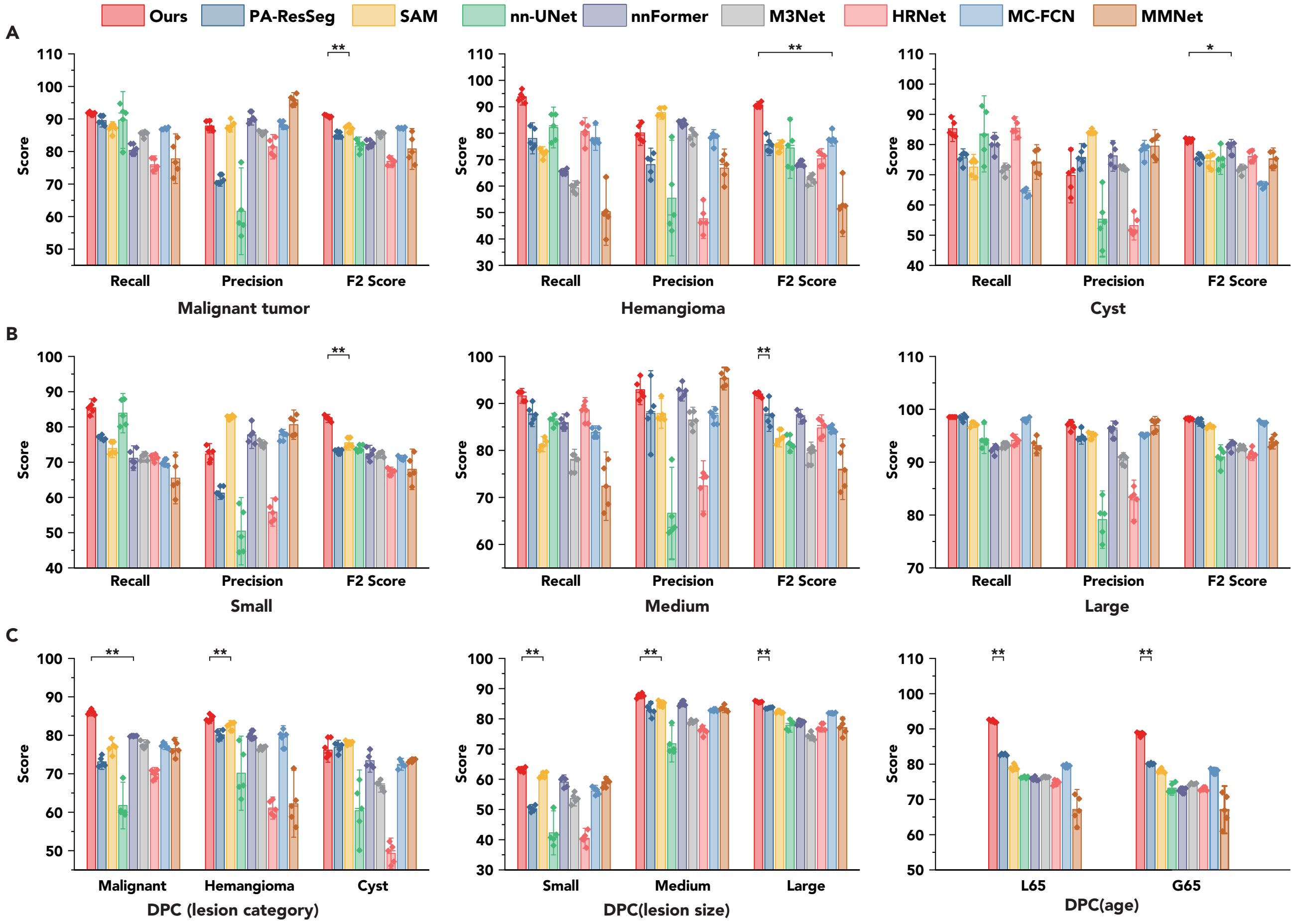


Figure 5. Our AI system outperformed existing approaches for liver lesion segmentation

Data are represented as mean ± SEM. We repeated the experiments five times, and the rhombus dots represent the scores of the test data.

(A) Comparison by various lesion categories. Compared to the best performance of competitive methods, our segmentation performance of liver lesions improved by 3.60%, 12.45%, and 5.86% on F2 score in the groups of malignant tumor, hemangioma, and cyst, respectively. This indicated that the comprehensive performance of our AI system in the three categories was superior to that of the baseline methods (** $p < 0.001$, two-sided t test; error bars are standard deviation).

(B) Comparison by various lesion sizes. According to the maximum diameter of lesions, we divided them into three levels: small, medium, and large, as shown in Table S1. MULLET showed remarkable enhancement of the performance in the group of small lesions, with improvements of 8.46% (compared with PA-ResSeg) and 6.92% (compared with SAM) in terms of recall and F2 score. In addition, compared with PA-ResSeg, MULLET also achieved improvements of 4.22% in the group of medium lesions. The ranges of all whiskers are standard deviations, and the coefficients are 1.5.

(C) Comparison of DPC score in terms of various lesion categories.

applicable AI-empowered system, MULLET, to assist radiologists in diagnosing liver lesions from multi-phase CECT scans. Extensive experiments on a large real-world clinical dataset that contains 1,229 CECT scans demonstrated the superior performance of MULLET. A further comparative study with experienced radiologists for liver lesion segmentation showed that MULLET analyzed multi-phase CECTs more efficiently and accurately. Without loss of generality, we also evaluated MULLET on two public datasets (details are provided in the STAR Methods section), which also achieved a promising performance compared with the state-of-the-art deep learning approaches for medical imaging segmentation.

The key innovations and superiority of MULLET over conventional methods benefit from the utilization of multi-phase contextual information in the misaligned regions of liver lesions. Intuitively, the input multi-phase images or extracted features by specific backbones from the multi-phase CECT scans can be simply concatenated.[25,26,31,36] However, the straightforward strategy cannot handle complex real-world scenarios such as the misalignment problem between phases. Although several deep learning methods[13,21,25,47] were designed by leveraging various feature filter modules to sift the beneficial features from other phases, the filter modules still suffer from losing useful cross-phase information due to the misalignment, especially for small lesions. Qu et al.[23] proposed a cross-phase non-local attention module to build the local alignment relationship between phases, but it relies heavily on inductive biases. Unlike the previous segmentation applications in medical image data analysis, which adopted a linear transform strategy (e.g., using concatenation or feature filtering) to fuse multi-phase contexts, MULLET can effectively utilize and aggregate multi-phase contextual information between phases. Specifically, we first introduced a

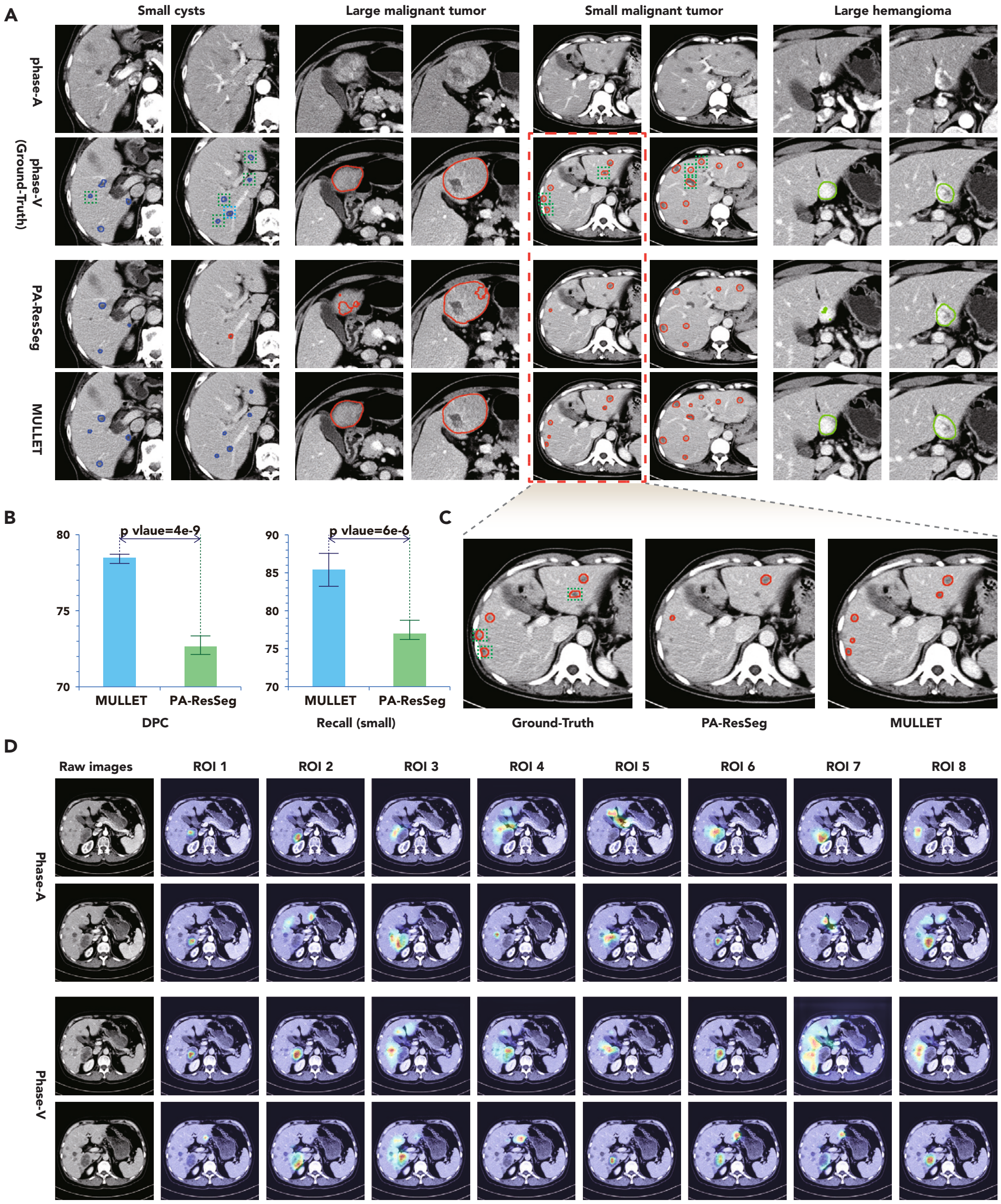


Figure 6. Visualization of liver lesion segmentation and ROIs maps

(A) Representative liver lesion segmentation results from multi-phase CECT images in the AIELDI dataset. Data are represented as mean ± SEM. The first and second rows are the ground-truth CECT segmentation in phase-A and phase-V. The two bottom rows are the segmentations produced by our AI system and PA-ResSeg (the best baseline), respectively. We present two adjacent CECT slices of each category, where the malignant tumors, hemangiomas, and cysts are marked by different colors. In the first case, our AI system perfectly detected all small cysts, while PA-ResSeg missed several small lesions (surrounded by green boxes). PA-ResSeg was able to detect a small lesion but predicted a wrong label (surrounded by blue boxes). In the third case, our AI system segmented five malignant tumors (surrounded by green boxes), while PA-ResSeg was not able to detect them. In the second and last cases, our AI system achieved more complete boundaries for large lesions than PA-ResSeg. These results further revealed that the segmentation produced by our AI system produced a much smoother and perfect outline compared to state-of-the-art methods.

(B) The performance comparison between MULLET and PA-ResSeg. MULLET significantly outperformed the best baseline PA-ResSeg.

(C) The magnified segmentation results of the small malignant tumor.

(D) The visualization of ROI's weight maps yielded from the multi-phase ROI encoder module. "Warm" filter colors (redder) highlight regions of high importance according to the ROI weight maps. Most ROI weight maps highlighted suspicious lesion regions, indicating the effectiveness of the proposed method. See also Figures S2–S4.

Table 4. Quantitative comparison between our method and state of the arts on the BRATS dataset

Methods	ET	TC	WT	AVG
MC-FCN	59.45±4.85	72.51±3.79	69.55±3.63	67.17±3.39
HRNet	63.31±3.61	69.85±4.04	68.40±5.25	67.18±4.13
M^{3}Net	64.99±1.76	68.52±4.77	70.73±3.16	68.08±1.63
MMNet	69.59±4.50	73.54±14.9	69.63±7.34	70.92±7.57
nnFormer	71.50±2.91	72.70±2.59	71.10±3.19	71.77±2.44
SAM	68.88±3.07	76.34±3.41	70.64±4.17	71.95±2.47
PA-ResSeg	70.70±5.46	76.81±5.33	73.95±3.05	73.82±4.13
nn-UNet	73.17±2.32	77.66±2.09	76.55±1.56	75.79±0.95
MULLET	**73.44±1.82**	**80.70±2.58**	**77.10±1.98**	**77.08±0.98**

Our AI system achieved the DPC of 73.44%, 80.70%, and 77.10% for ET, TC, and WT, respectively. Compared with the best baseline nn-UNet, the average DPC improvement is 1.29% (p value = 2.9e-4).

multi-phase ROI encoder, which produces several important ROIs for each CT slice under the supervision of the cross-phase global and local features. Then, we applied a transformer to explore the inter-phase and intra-phase relationship between the ROIs of different phases, respectively. Consequently, compared with conventional methods using feature filtering and inductive biases, our method can better learn and utilize multi-phase contexts in the misaligned locations. It is worth mentioning that MULLET also enabled efficient dealing with small lesions, which is of great importance to radiologists in clinical diagnosis.

Limitations of the study

Although our AI system achieved promising performance for liver lesion segmentation, some limitations are still not well addressed in MULLET. (i) It is still challenging to yield superb segmentation for very small lesions due to the limited thickness (3.0–5.0 mm) of CECT images, despite our system having achieved promising results compared with state of the arts for very small lesions. Note that MULLET was able to capture multi-phase contextual information for small lesions better. In contrast, some specific modules dedicated to segmenting very small lesions (e.g., ≤ 1 cm) need to be designed. (ii) It is valuable to leverage auxiliary modality data along with the multi-phase CECT images to improve model performance further. It has been proved that other modality data collected with the CECT scans may assist the diagnosis of liver lesions.[13,14,23,31,47]

In future work, we will collect some abdominal CECT scans with smaller thicknesses and explore small object detection technologies in our model, e.g., multi-scale features. This will lead to a more accurate AI system for fine-grained lesion segmentation. In addition, we will continue to enhance MULLET by leveraging multi-modal data as auxiliary information to improve the segmentation performance further. Furthermore, we plan to build a more user-friendly diagnostic system based on our MULLET model for more general medical imaging segmentation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

Table 5. Quantitative comparison between our method and state of the arts in terms of segmentation of abdominal organs on the CHAOS (MRI only) dataset

Methods	liver	right kidney	left kidney	spleen	AVG
MC-FCN	84.16±2.70	64.56±8.04	59.27±5.86	57.95±15.55	66.49±4.04
M^{3}Net	80.13±1.76	67.43±4.77	65.23±3.16	61.53±1.63	68.58±1.63
MMNet	87.10±6.57	74.21±12.46	63.54±24.47	55.31±33.16	70.04±13.25
SAM	83.86±5.27	70.83±11.05	67.76±12.12	65.89±13.73	72.09±4.96
PA-ResSeg	89.24±9.37	74.50±10.11	73.98±10.39	67.55±26.19	76.32±7.46
nnFormer	88.82±11.98	75.66±11.99	70.06±12.56	65.68±15.78	74.96±10.06
HRNet	88.74±12.49	78.58±12.18	71.00±15.26	69.27±18.04	76.90±8.37
nn-Unet	92.53±2.64	78.23±9.90	76.04±15.68	71.25±14.69	79.51±7.15
MULLET	91.37±1.69	**85.06±4.72**	**83.68±9.97**	**79.23±7.4**	**84.84±5.02**

Our AI system significantly improved the segmentation of the right kidney, left kidney, and spleen, i.e., an improvement of 6.48%, 7.64%, and 7.98%, respectively.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108183.

ACKNOWLEDGMENTS

This work was supported by the Key Research and Development Program of Zhejiang Province (No. 2021C03121), the National Natural Science Foundation of China (Grant Nos. 62202422, 61972349, and 82072650), the Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province (2020E10004), Central Guidance on Local Science, Technology Development Fund of Zhejiang Province (2023ZY1017), and Zhejiang University-Pujian Technology Joint Research Center.

AUTHOR CONTRIBUTIONS

H.W., Y.D., and J.B. conceived the project. L.W. and H.W. developed the MULLET model and conducted the computational experiments. Y.C., Z.S., and Y.D. collected the multi-phase CECT from the hospital. L.W., H.W., and X.Z. drafted, revised, and edited the manuscript. T.Z., N.S., and G.T. provided additional pathological insights into the experimental results. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors have no conflicts of interest to declare.

Received: April 14, 2023
Revised: July 22, 2023
Accepted: October 9, 2023
Published: October 13, 2023

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
2. Kelley, R.K., and Greten, T.F. (2021). Hepatocellular carcinoma — origins and outcomes. N. Engl. J. Med. *385*, 280–282. https://doi.org/10.1056/NEJMcibr2106594.
3. Llovet, J.M., Kelley, R.K., Villanueva, A., Singal, A.G., Pikarsky, E., Roayaie, S., Lencioni, R., Koike, K., Zucman-Rossi, J., and Finn, R.S. (2021). Hepatocellular carcinoma. Nat. Rev. Dis. Prim. *7*, 6. https://doi.org/10.1038/s41572-020-00240-3.
4. Mitchell, D.G., Bruix, J., Sherman, M., and Sirlin, C.B. (2015). Li-rads (liver imaging reporting and data system): Summary, discussion, and consensus of the li-rads management working group and future directions. Hepatology *61*, 1056–1065. https://doi.org/10.1002/hep.27304.
5. Mitsuzaki, K., Yamashita, Y., Ogata, I., Nishiharu, T., Urata, J., and Takahashi, M. (1996). Multiple-phase helical ct of the liver for detecting small hepatomas in patients with liver cirrhosis: contrast-injection protocol and optimal timing. Am. J. Roentgenol. *167*, 753–757. https://doi.org/10.2214/ajr.167.3.8751695.
6. Sangiovanni, A., Manini, M.A., Iavarone, M., Romeo, R., Forzenigo, L.V., Fraquelli, M., Massironi, S., Della Corte, C., Ronchi, G., Rumi, M.G., et al. (2010). The diagnostic and economic impact of contrast imaging techniques in the diagnosis of small hepatocellular carcinoma in cirrhosis. Gut *59*, 638–644. https://doi.org/10.1136/gut.2009.187286.
7. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al. (2022). A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. Nat. Commun. *13*, 2096–2111. https://doi.org/10.1038/s41467-022-29637-2.
8. Burrows, L., Chen, K., Guo, W., Hossack, M., McWilliams, R.G., and Torella, F. (2022). Evaluation of a hybrid pipeline for automated segmentation of solid lesions based on mathematical algorithms and deep learning. Sci. Rep. *12*, 14216–14311. https://doi.org/10.1038/s41598-022-18173-0.
9. Lösel, P.D., van de Kamp, T., Jayme, A., Ershov, A., Faragó, T., Pichler, O., Tan Jerome, N., Aadepu, N., Bremer, S., Chilingaryan, S.A., et al. (2020). Introducing biomedisa as an open-source online platform for biomedical image segmentation. Nat. Commun. *11*, 5577. https://doi.org/10.1038/s41467-020-19303-w.
10. Primakov, S.P., Ibrahim, A., van Timmeren, J.E., Wu, G., Keek, S.A., Beuque, M., Granzier, R.W.Y., Lavrova, E., Scrivener, M., Sanduleanu, S., et al. (2022). Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nat. Commun. *13*, 3423. https://doi.org/10.1038/s41467-022-30841-3.
11. Fu, F., Wei, J., Zhang, M., Yu, F., Xiao, Y., Rong, D., Shan, Y., Li, Y., Zhao, C., Liao, F., et al. (2020). Rapid vessel segmentation and reconstruction of head and neck angiograms using 3d convolutional neural network. Nat. Commun. *11*, 4829. https://doi.org/10.1038/s41467-020-18606-2.
12. Moghbel, M., Mashohor, S., Mahmud, R., and Saripan, M.I.B. (2018). Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. Artif. Intell. Rev. *50*, 497–537. https://doi.org/10.1007/s10462-017-9550-x.

13. Xu, Y., Cai, M., Lin, L., Zhang, Y., Hu, H., Peng, Z., Zhang, Q., Chen, Q., Mao, X., Iwamoto, Y., et al. (2021). Pa-resseg: A phase attention residual network for liver tumor segmentation from multiphase ct images. Med. Phys. *48*, 3752–3766. https://doi.org/10.1002/mp.14922.
14. Zhang, Y., Peng, C., Peng, L., Huang, H., Tong, R., Lin, L., Li, J., Chen, Y.W., Chen, Q., Hu, H., et al. (2021). Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (Springer). https://doi.org/10.1007/978-3-030-87193-2_7.
15. Wang, X., Han, S., Chen, Y., Gao, D., and Vasconcelos, N. (2019). Volumetric attention for 3d medical image segmentation and detection. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019 (Springer). https://doi.org/10.1007/978-3-030-32226-7_20.
16. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., and Heng, P.A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Trans. Med. Imag. *37*, 2663–2674. https://doi.org/10.1109/TMI.2018.2845918.
17. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al. (2023). The liver tumor segmentation benchmark (lits). Med. Image Anal. *84*, 102680. https://doi.org/10.1016/j.media.2022.102680.
18. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods *18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z.
19. Roberts, L.R., Sirlin, C.B., Zaiem, F., Almasri, J., Prokop, L.J., Heimbach, J.K., Murad, M.H., and Mohammed, K. (2018). Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. Hepatology *67*, 401–421. https://doi.org/10.1002/hep.29487.
20. Ayuso, C., Rimola, J., Vilana, R., Burrel, M., Darnell, A., García-Criado, Á., Bianchi, L., Belmonte, E., Caparroz, C., Barrufet, M., et al. (2018). Diagnosis and staging of hepatocellular carcinoma (hcc): current guidelines. Eur. J. Radiol. *101*, 72–81. https://doi.org/10.1016/j.ejrad.2018.01.025.
21. Jiang, X., Luo, Q., Wang, Z., Mei, T., Wen, Y., Li, X., Cheng, K.T., and Yang, X. (2020). Multi-phase and multi-level selective feature fusion for automated pancreas segmentation from ct images. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020 (Springer). https://doi.org/10.1007/978-3-030-59719-1_45.
22. Ruskó, L., Bekes, G., and Fidrich, M. (2009). Automatic segmentation of the liver from multi-and single-phase contrast-enhanced ct images. Med. Image Anal. *13*, 871–882. https://doi.org/10.1016/j.media.2009.07.009.
23. Qu, T., Wang, X., Fang, C., Mao, L., Li, J., Li, P., Qu, J., Li, X., Xue, H., Yu, Y., and Jin, Z. (2022). M^3Net: A multi-scale multi-view framework for multi-phase pancreas segmentation based on cross-phase non-local attention. Med. Image Anal. *75*, 102232. https://doi.org/10.1016/j.media.2021.102232.
24. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Association for Computational Linguistics-ACL 2019 (ACL). https://doi.org/10.18653/v1/P19-1656.
25. Sun, C., Guo, S., Zhang, H., Li, J., Chen, M., Ma, S., Jin, L., Liu, X., Li, X., and Qian, X. (2017). Automatic segmentation of liver tumors from multiphase contrast-enhanced ct images based on fcns. Artif. Intell. Med. *83*, 58–66. https://doi.org/10.1016/j.artmed.2017.03.008.
26. Ouhmich, F., Agnus, V., Noblet, V., Heitz, F., and Pessaux, P. (2019). Liver tissue segmentation in multiphase ct scans using cascaded convolutional neural networks. Int. J. Comput. Ass. Rad. *14*, 1275–1284. https://doi.org/10.1007/s11548-019-01989-z.
27. Wang, M., Fu, F., Zheng, B., Bai, Y., Wu, Q., Wu, J., Sun, L., Liu, Q., Liu, M., Yang, Y., et al. (2021). Development of an ai system for accurately diagnose hepatocellular carcinoma from computed tomography imaging data. Br. J. Cancer *125*, 1111–1121. https://doi.org/10.1038/s41416-021-01511-w.
28. Senthilkumaran, N., and Vaithegi, S. (2016). Image segmentation by using thresholding techniques for medical images. Comput. Syst. Sci. Eng. *6*, 1–13. https://doi.org/10.5121/cseij.2016.6101.
29. Raja, N.S.M., Fernandes, S.L., Dey, N., Satapathy, S.C., and Rajinikanth, V. (2018). Contrast enhanced medical mri evaluation using tsallis entropy and region growing segmentation. J. Ambient Intell. Hum. Comput. *1*. https://doi.org/10.1007/s12652-018-0854-8.
30. Wang, H., Li, L., Chi, L., and Zhao, Z. (2019). Autism screening using deep embedding representation. In Computational Science–ICCS 2019 (Springer). https://doi.org/10.1007/978-3-030-22741-8_12.
31. Hasegawa, R., Iwamoto, Y., Lin, L., Hu, H., and Chen, Y.W. (2020). Automatic segmentation of liver tumor in multiphase ct images by mask r-cnn. In Conference on Life Sciences and Technologies-LifeTech 2020 (IEEE). https://doi.org/10.1109/LifeTech48969.2020.1570619011.
32. Qin, D., Wang, H., Liu, Z., Xu, H., Zhou, S., and Bu, J. (2022). Hilbert distillation for cross-dimensionality networks. In Neural Information Processing Systems-NeurIPS 2022) (Curran Associates Inc). https://doi.org/10.48550/arXiv.2211.04031.
33. Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A.B., and Kronzer, A. (2018). Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM Trans. Comput. Biol. Bioinform. *15*, 1968–1978. https://doi.org/10.1109/TCBB.2018.2827029.
34. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision-ECCV 2018 (Springer). https://doi.org/10.1007/978-3-030-01234-2_49.
35. Yan, K., Bagheri, M., and Summers, R.M. (2018). 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In Medical Image Computing and Computer Assisted Intervention-MICCAI 2018 (Springer). https://doi.org/10.1007/978-3-030-01234-2_49.
36. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., and Ben Ayed, I. (2019). Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. IEEE Trans. Med. Imag. *38*, 1116–1126. https://doi.org/10.1109/TMI.2018.2878669.
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Neural Information Processing Systems-NeurIPS 2017 (Curran Associates Inc.). https://doi.org/10.5555/3295222.3295313.
38. Mansoori, B., Erhard, K.K., and Sunshine, J.L. (2012). Picture archiving and communication system (pacs) implementation, integration & benefits in an integrated health system. Acad. Radiol. *19*, 229–235. https://doi.org/10.1016/j.acra.2011.11.009.
39. Onken, M., Eichelberg, M., Riesmeier, J., and Jensch, P. (2011). Digital imaging and communications in medicine. In Biomedical Image Processing, T.M. Deserno, ed. (Springer), pp. 427–454. https://doi.org/10.1007/978-3-642-15816-2_17.
40. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Computer Vision and Pattern Recognition-CVPR 2016 (IEEE). https://doi.org/10.1109/CVPR.2016.90.
41. Wang, H., and Avillach, P. (2021). Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning. JMIR Med. Inf. *9*, e24754. https://doi.org/10.2196/24754.
42. Ryoo, M., Piergiovanni, A., Arnab, A., Dehghani, M., and Angelova, A. (2021). Tokenlearner: Adaptive space-time tokenization for videos. In Neural Information Processing Systems-NeurIPS 2021 (Curran Associates Inc.).
43. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., and Yu, Y. (2023). nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Trans. Image Process. *32*, 4036–4045. https://doi.org/10.1109/TIP.2023.3293771.
44. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2021). Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. *43*, 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.
45. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imag. *34*, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.
46. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al. (2021). Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. *69*, 101950. https://doi.org/10.1016/j.media.2020.101950.
47. Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., and He, Z. (2021). Modality-aware mutual learning for multi-modal medical image segmentation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2021 (Springer). https://doi.org/10.1007/978-3-030-87193-2_56.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Selected CT images	This paper	https://github.com/shenhai1895/Multi-phase-Liver-Lesion-Segmentation
CHAOS	https://doi.org/10.1016/j.media.2020.101950	2021
BraTS	https://www.med.upenn.edu/cbica/brats2020/	2020
Software and algorithms		
Python	https://www.python.org/	Version 3.9
numpy	https://numpy.org/	1.24.3
scikit-image	https://scikit-image.org/	0.19.3
SimpleITK	https://simpleitk.org/	2.2.1
segmentation_models_pytorch	https://github.com/qubvel/segmentation_models.pytorch	0.3.2
pytorch	https://pytorch.org/	2.0.1
torchvision	https://github.com/pytorch/vision	0.15.2
Code for liver lesion segmentation	This paper	https://github.com/shenhai1895/Multi-phase-Liver-Lesion-Segmentation

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Haishuai Wang(haishuai.wang@zju.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The authors released partial data (i.e., 14 raw multi-phase CECT scans collected from the Second Affiliated Hospital, Zhejiang University School of Medicine) with permission to support the results in this study, which are available at Google Drive. Due to the regulations and privacy policies, the full datasets remain under custody of hospitals. The BraTS and CHAOS benchmarking datasets that utilized in this study originated from public data repositories, which are available at https://www.med.upenn.edu/cbica/brats2020/ and https://chaos.grand-challenge.org/, respectively.
- Our AI system is open source, and the source code to implement MULLET is publicly available on GitHub (https://github.com/shenhai1895/Multi-phase-Liver-Lesion-Segmentation).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The overall architecture of MULLET is illustrated in Figure 1, which consists of four key components: (1) a CNN-based feature extraction module to extract features from all slices of all phases; (2) a multi-phase ROI encoder module to generate several ROIs for each CECT slice according to its cross-phase features and embed the ROIs into features; (3) a Transformer-based multi-phase context exploration module to explore the inter-phase and intra-phase relationship between the ROIs of different phases; and (4) a decoder to produce the segmentation as the output. The detailed model architectures are provided in supplemental information.

Dataset pre-processing

The AIELDI dataset was used for the AI model development. MULLET takes the multi-phase CECT images in the AIELDI dataset as the input. We preprocessed the acquired CECTs before feeding an image into the model to improve computational efficiency. Multi-phase CECT data normalization was the first step of image preprocessing. Due to the various physical resolution (i.e., ranging from 0.472 mm to 0.890 mm) of the acquired multi-phase CECT images, we normalized all images to an isotropic resolution of 0.7×0.7×5.0 mm^3. For example, if the resolution is higher than 0.7 mm, down-sampling was employed; otherwise, up-sampling was applied to the multi-phase CECT images. Moreover, to reduce the effect of extreme values, especially in the area of metal artifacts, we clipped the intensity values of each multi-phase CECT scan to [-55, 155] before intensity normalization. Additionally, following the standard protocol of image processing in deep learning, the voxel-wise intensities V were normalized to the interval [-1, 1], such that

$$V_{norm} = \frac{2 * (V - V_{min})}{V_{max} - V_{min}} - 1. \quad \text{(Equation 1)}$$

Feature extractor

Unlike PA-ResSeg[13] and SAM,[14] we sampled a sequence of CECT images as the input to explore multi-phase contexts along z-axis. As shown in Figure 1, ResNet-34 was directly employed to extract distinct features from the CECTs. To reduce the burden of computing resources, we randomly cropped N slices with a size of 256×256 from the CECT images to be fed into the model and obtained the features with a size of 16 × 16. In this step, the CECT images from different phases were fed into a CNN backbone to extract the features of the two phases. Let $X_a = \{X_a^1, \cdots, X_a^N\}, X_v = \{X_v^1, \cdots, X_v^N\}$ be the image sequences in phase-A and phase-V, where $X_a^i|_{i=1}^{N} \in \mathbb{R}^{H \times W}$ and $X_v^i|_{i=1}^{N} \in \mathbb{R}^{H \times W}$. The N denotes the number of CECT slices in each phase. We set $N = 9$ following VA-Mask RCNN.[15] The H and W denote the height and width of each CECT image. A pretrained CNN network (e.g., ResNet34) was applied to extract high-level features $\mathcal{F}_a = \{\mathcal{F}_a^1, \cdots, \mathcal{F}_a^N\}$ and $\mathcal{F}_v = \{\mathcal{F}_v^1, \cdots, \mathcal{F}_v^N\}$ from X_a and X_v, as follows:

$$\mathcal{F}_a^i = \text{CNN}\left(X_a^i\right); \quad \text{(Equation 2)}$$

$$\mathcal{F}_v^i = \text{CNN}\left(X_v^i\right), \quad \text{(Equation 3)}$$

where $\mathcal{F}_a, \mathcal{F}_v \in \mathbb{R}^{c \times N \times h \times w}$, $h = \frac{H}{16}$, $w = \frac{W}{16}$. c is the number of the channels.

Multi-phase ROI encoder

After extracting the features, we applied our proposed multi-phase ROI encoder to produce several ROIs for each CECT slice, as shown in Figure 2. The produced ROIs are mainly concentrated in liver lesion regions. We take phase-V as an example to introduce the details of this module because phase-A was handled in the same way. In this module, we leveraged the cross-phase features extracted from the auxiliary phase (i.e., phase-A) to produce precise ROIs for the target phase (i.e., phase-V). This module first adopted attention mechanisms to extract cross-phase global and local features for all phases mutually (Figure 3), which enabled this module to yield more precise ROIs. Then, thanks to the extracted cross-phase features, we applied a spatial attention module on multi-phase contexts to produce several ROI weight maps and encoded these maps into features by global pooling. Specifically, cross-phase global features were produced along two different axes. The global information along the z-axis was extracted by a Cross-phase Global Inter-slice Attention (CGIA) module. The CGIA squeezed the features of all phases along the z-axis and derived the weights between each slice in one phase and all slices in the other phase. Similarly, a Cross-phase Global Spatial Attention (CGSA) module was adopted to squeeze the features along the axial plane and explore the weights of each feature point between phases. A Cross-phase Local Attention (CLA) module was then used to aggregate the corresponding neighboring information for each feature point through 3D convolutional layers. Thus, the cross-phase feature representation $\mathcal{F}_{c,v}$ can be represented by $\mathcal{F}_{c,v} = \text{CGIA}(\mathcal{F}_v, \mathcal{F}_a) + \text{CGSA}(\mathcal{F}_v, \mathcal{F}_a) + \text{CLA}(\mathcal{F}_v, \mathcal{F}_a)$. The three attention mechanisms enabled aggregating different cross-phase information from three aspects (i.e., the global information along the z-axis, the global information along the axial plane, and pixel-wise features from phase-A). The summation of the cross-phase features $\mathcal{F}_{c,v} \in \mathbb{R}^{c \times N \times h \times w}$ and the original features (i.e., $\mathcal{F}_v$) were considered as multi-phase features $\mathcal{F}_{m,v} = \{\mathcal{F}_{m,v}^1, \cdots, \mathcal{F}_{m,v}^N\}$, followed by a batch normalization layer and a convolution layer.

For the multi-phase features $\mathcal{F}_{m,v}^i$ of the $i-$th slice, we utilized a spatial attention module $\varphi(\mathcal{F}_{m,v}^i)$ to generate S spatial weighted maps with the size of $h \times w$. The spatial attention module $\varphi(\mathcal{F}_{m,v}^i)$ contains multiple Conv2D layers, in which the *out_channel* of the last Conv2D layer is set to S and the size of its output is $S \times h \times w$. We first applied a softmax layer for each map to derive the weights. Then, we use the S spatial weighted maps to encode $\mathcal{F}_{m,v}^i$ to S ROI features by a weighted sum. Therefore, the encoded features of ROIs $\mathcal{F}_{ROI,v}^i$ for $i-$th CECT image in the input data is given by

$$\mathcal{F}_{ROI,v}^i = \rho\left(\mathcal{F}_{m,v}^i \odot \gamma\left(\varphi\left(\mathcal{F}_{m,v}^i\right)\right)\right) = \rho\left(\mathcal{F}_{m,v}^i \odot \gamma\left(\sigma\left(\text{Conv2D}\left(\mathcal{F}_{m,v}^i\right)\right)\right)\right) \quad \text{(Equation 4)}$$

where $\mathcal{F}_{ROI,v}^i \in \mathbb{R}^{c \times S}$ $\odot$ represents element-wise multiplication, $\gamma(\cdot)$ denotes a broadcasting function, and $\sigma(\cdot)$ is a softmax function. Spatial global sum pooling $\rho(\cdot)$ was applied on top of them to reduce the dimensionality. We applied the same process to generate $\mathcal{F}_{ROI,a}^i$ for each slice in phase A.

Cross-phase local attention

The Cross-phase Local Attention (CLA) aims at filtering pixel-wise local information from the auxiliary phase. A local representation for $\mathcal{F}_v$ was implemented by a 3×3×3 convolution layer, which aggregated neighboring information of each pixel. We performed an element-wise multiplication between $\mathcal{F}_a$ and the local representation and then applied a Linear layer to reduce the channel into 1. A *sigmoid* function was applied to calculate the filter maps with the dimension of $1 \times N \times h \times w$.

Finally, we can obtain the cross-phase local features for phase-V through the element-wise multiplication between the filter maps and the local features of phase-A.

Cross-phase Global Inter-slice attention

The Cross-phase Global Inter-slice Attention (CGIA) explored the global contexts along the $z-$ axis from the auxiliary phase. Specifically, according to the architecture of the standard Transformer, a *query* representation for $\mathcal{F}_v$ was implemented by a 1×1 convolution layer followed by a batch normalization layer. Meanwhile, two 1×1 convolution layers and batch normalization layers were used to derive a *key* and *value* representations from $\mathcal{F}_a$, respectively. Then, spatial global average pooling was applied on the *query* and *key* to reduce the dimensionality to $c \times N$. We performed a matrix multiplication between the reduced *query* and the transpose of the *key* and applied a *softmax* layer to calculate the inter-slice attention map.

After that, we performed a matrix multiplication between the inter-phase attention map and the transpose of the *value* representation to yield the cross-phase global features.

Cross-phase Global Spatial Attention

The Cross-phase Global Spatial Attention (CGSA) aims to explore the global spatial contexts from the auxiliary phase. A *query* representation for $\mathcal{F}_v$ was implemented by a 1×1 convolution layer followed by a batch normalization layer. Meanwhile, two 1×1 convolution layers and batch normalization layers were adopted to derive *key* and *value* representations from $\mathcal{F}_a$, respectively. In particular, a mean function was applied on *query* and *key* along $z-$ axis to reduce the dimensionality. We performed a matrix multiplication between the reduced *query* and the transpose of the reduced *key* and applied a *softmax* layer to calculate the inter-phase attention map. We also performed a matrix multiplication between the inter-phase attention map and the transpose of the *value* to get the cross-phase global spatial features.

Multi-phase context exploration

We first explored the relationship (i.e., multi-phase contexts) between the ROIs of all phases via a Transformer-based module, aiming to alleviate the effects of misalignment across multiple phases without resorting to any inductive biases. The Transformer consists of two multi-head attention modules. The first multi-head attention module aims at learning the intra-phase contexts, which represent the relationship between the ROIs of one phase, while the second multi-head attention module was used to explore the relationship between the ROIs of other phases (i.e., inter-phase contexts), donated as

$$\mathcal{F}_{intra,v} = MHA(\mathcal{F}_{ROI,v}, \mathcal{F}_{ROI,v}, \mathcal{F}_{ROI,v}) + \mathcal{F}_{ROI,v}; \qquad \text{(Equation 5)}$$

$$\mathcal{F}_{inter,v} = MHA(\mathcal{F}_{intra,v}, \mathcal{F}_{ROI,v}, \mathcal{F}_{ROI,v}) + \mathcal{F}_{intra,v} \qquad \text{(Equation 6)}$$

Note that the multi-phase contexts were achieved through the residual connection of intra-phase and inter-phase contexts. The ROIs embedding with multi-phase contextual information (with the shape of $c \times S$) will be used to generate the final segmentation. However, the shape (i.e., $c \times S$) is not matched with the input size of the decoder. Therefore, it is necessary to remap the ROIs embedding with multi-phase contextual information (with the shape of $c \times S$) into the original size (of $c \times h \times w$) through an inverse operation on the ROIs embedding. We adopted a TokenFuser[42] module as the ROIFuser module to fuse the multi-phase contexts and remap it back to its original spatial resolution of $\mathcal{F}_v^i$, donated as

$$\widehat{\mathcal{F}}_v^i = ROIFuser\left(\left(FFN\left(\mathcal{F}_{inter,v}^i + \mathcal{F}_{intra,v}^i\right)\right), \mathcal{F}_v^i\right) + \mathcal{F}_v^i. \qquad \text{(Equation 7)}$$

where *FFN* is a feedforward network, and $\widehat{\mathcal{F}}_v^i \in \mathbb{R}^{c \times h \times w}$. The ROIFuser module was designed to solve this challenge rather than solving the misalignment between phases.

Segmentation decoder

In the final step, we applied DeepLabV3 plus[34] as the decoder on $\widehat{\mathcal{F}}_v$ to produce the segmentation Seg_v^i of $i-$ th CECT slice in Phase-V:

$$Seg_v^i = decoder\left(\widehat{\mathcal{F}}_v^i\right) \qquad \text{(Equation 8)}$$

where $Seg_v^i \in \mathbb{R}^{C \times H \times W}$ and C is the number of liver lesion categories. Each pixel was labeled as one of the categories in the segmentation, i.e., background, malignant tumor, hemangioma, and cyst. When the pixels of a lesion were labeled as different categories, we took the majority as the lesion's label.

Training details and evaluation metrics

Training details

The framework was implemented via PyTorch library, using the Adam optimizer to minimize the loss functions and to optimize network parameters by backpropagation. The loss function is the sum of a dice loss and a cross-entropy loss. A learning rate of 0.0001 and a mini-batch size of 10 were used in the segmentation network. We trained our model for 30 epochs in the liver lesion segmentation task, taking about 10 hours. All deep neural networks were trained with 4 Nvidia GeForce RTX 3090 GPUs.

Evaluation metrics

To evaluate segmentation performance, we adopted Dice Per Case (DPC) and Volumetric Overlap Error (VOE) to measure the accuracy of semantic segmentation for multi-phase CECT images. In addition, recall, precision, and F2-score were calculated to validate the classification performance after the segmentation. DPC represents the spatial overlap between ground truth and segmentation, and VOE denotes the error rate of the segmentation. The F2 score represents a comprehensive measure of recall and precision. The evaluation metrics are formally defined in Equation 9.

$$DPC = \frac{2|R \cap G|}{|R| + |G|}, \qquad VOE = 1 - \frac{R \cap G}{R \cup G}$$

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP} \qquad , \qquad \text{(Equation 9)}$$

$$F2 - score = \frac{5 * Recall * Precision}{Recall + 4 * Precision}$$

where *R*, *G*, TP, FP, and FN denote the segmentation result, the ground-truth, true positives, false positives, and false negatives, respectively. Note that it is the lower the better for VOE, while it is higher the better for other metrics.

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise stated in the figure legends, data is shown as mean ± SEM. Statistical comparisons between the two groups were evaluated by a two-tailed t-test using Excel. Performance of the lesion segmentation networks is assessed by DPC, VOE, recall, precision, and F2 score, as shown in Figure 5 and Tables 2, 3, 4, and 5. The MULLET was implemented using Pytorch, Torchvision, and Python. See key resources table for additional details.