MCAF-Net: A Multi-Feature Complementary and Adaptive Fusion Network for Automatic Liver Segmentation in Computed Tomography Images

Authors: Shen Tong, Wang Jianlin, Runqiu Gu, Wang Yadi

Date: 2025-06-22T00:00:00+00:00

Abstract

To more accurately assist physicians in diagnosing liver diseases and planning surgical procedures, accurate and stable automatic liver segmentation from CT images is an urgent problem that needs to be addressed. However, medical images inevitably contain noise and artifacts, which can easily influence segmentation algorithms. To tackle this issue, this paper proposes MCAF-Net, which embeds MCCA into the bottleneck layer to generate rich and complete feature representations through multi-feature complementarity, thereby reducing information loss and mitigating the impact of noise and artifacts via feature interaction. Moreover, to more accurately identify liver boundaries, the encoder and decoder are connected through AMFM, enhancing the perception of contextual and multi-scale information, thus enabling fine-grained segmentation of liver edges. Experimental results on the LiTS2017 and noisy LDCT datasets demonstrate that MCAF-Net outperforms other mainstream algorithms in reducing the effects of noise and artifacts as well as in liver edge identification, achieving DSC and Jarracd scores of 96.24% and 92.83%, respectively, on the LiTS2017 dataset. The results on the LDCT dataset further indicate that MCAF-Net exhibits certain robustness and noise-resistant performance.

Full Text

Preamble

Vol. XX, No. X, XXX 20XX
NUCLEAR TECHNIQUES

MCAF-Net: Multi-Feature Complementary and Adaptive Fusion Network for Automatic Liver Segmentation in CT Images

Shen Tong[1], Wang Jianlin[1], Gu Runqiu[2], Wang Yadi[1]
[1] Southwest University of Science and Technology, Mianyang 621010, China
[2] Mianyang Central Hospital, Mianyang 621010, China

Abstract

[**Background**] Medical images inevitably contain noise and artifacts generated during the imaging process, which can adversely affect segmentation algorithms. Accurate and robust automatic liver segmentation from Computed Tomography (CT) images is essential for assisting physicians in diagnosing liver diseases and planning surgical procedures. [**Purpose**] This study aims to develop a novel image segmentation algorithm for precise liver region extraction from CT images. [**Methods**] We propose the Multi-feature Complementary and Adaptive Fusion Network (MCAF-Net), which embeds Multi-feature Complementary Cross-Attention (MCCA) into the bottleneck layer. MCCA generates rich and complete feature representations through multi-feature complementation to reduce information loss and employs feature interaction via cross-attention to mitigate the influence of noise and artifacts. Additionally, to achieve more precise liver localization, the encoder and decoder are connected through an Adaptive Multi-Scale Feature-fusion Module (AMFM), which enhances perception of contextual and multi-scale information, thereby enabling fine-grained segmentation of liver boundaries. [**Results**] Experimental results on the LiTS2017 dataset and the noisy LDCT dataset demonstrate that MCAF-Net outperforms other mainstream algorithms in reducing noise and artifact effects while recognizing liver edges. On LiTS2017, the Dice Similarity Coefficient (DSC) and Jaccard index reach 96.24% and 92.83%, respectively. Results on the LDCT dataset indicate that MCAF-Net possesses robust anti-noise capabilities. [**Conclusions**] The proposed MCAF-Net effectively addresses the challenges of noise and artifacts in CT images while achieving state-of-the-art liver segmentation performance.

Keywords: Liver segmentation; Multi-feature complementary; Adaptive; Anti-noise performance
Classification: TN957.52

DOI: 10.11889/j.0253-3219.2025.hjs.48.240548
CSTR: 32193.14.hjs.CN31-1342/TL.2025.48.240548

---

1 Introduction

The liver is one of the most vital human organs with multiple critical functions. Liver segmentation from abdominal medical images constitutes a key prerequisite for surgical planning, liver cancer detection, and interventional therapy. Traditional segmentation algorithms typically rely solely on local pixel or neighborhood information, lacking understanding of global contextual information, and thus cannot effectively segment complex structures with morphological

variations or irregularly shaped tissues [?]. Deep learning algorithms capture contextual semantic information in images through stacked convolutional layers, and with the rapid advancement of deep learning, Convolutional Neural Network (CNN)-based medical image segmentation has gradually replaced traditional methods as a focal point of research [?].

CNN architectures have provided valuable insights for medical image segmentation tasks. Long et al. [?] transformed fully connected layers in traditional CNNs into convolutional layers, creating Fully Convolutional Networks (FCN), which introduced new perspectives for subsequent semantic segmentation research. Ronneberger et al. [?] proposed U-Net with encoder-decoder architecture, which has been widely adopted and improved upon by many scholars, including U-Net++ [?] and SAR-U-Net [?].

Medical images are susceptible to variations in tissue size and noise, requiring segmentation algorithms to address issues of noise and information inconsistency. Specifically, CT image segmentation faces two major challenges: (1) CT images inevitably contain noise due to imaging equipment and procedures [?]; and (2) size variations in tissues or lesions cause imaging inconsistencies, while patient motion results in blurred organ edges and other interference factors. Noise and artifacts in CT images may cause algorithms to misclassify non-liver tissues with contrast similar to liver tissue or with blurred boundaries, leading to over-segmentation [?]. To address this, Trans-UNet [?] and Swin-UNet [?] simulate global contextual information by computing correlations between positions, thereby identifying and ignoring artifacts to some extent. However, when input data contains noise, distinguishing between noise and useful features becomes difficult, potentially leading to incorrect contextual correlations [?]. Compared to relying solely on global context, multi-feature fusion considers diverse features, enabling more comprehensive image analysis and reducing noise impact on segmentation results. Li et al. [?] proposed MFA-Net, which utilizes multi-scale associations to provide diverse contextual information for effective lesion perception, particularly for lesions of varying sizes with blurred edges. Wang et al. [?] proposed UC-TransNet, which establishes global context models through multi-scale feature aggregation, reducing local artifact effects to some degree. Xie et al. [?] proposed SegFormer, which demonstrates strong robustness against local artifacts thanks to its Transformer architecture and multi-scale fusion strategy. However, the multi-scale feature fusion strategies in SegFormer, UC-TransNet, and MFA-Net may cause information loss and insufficient precision during pixel-level segmentation due to limited sampling rates. Furthermore, feature fusion between encoder and decoder can strengthen information flow from shallow to deep layers, effectively preserving features and enabling gradient backpropagation. Simple addition or concatenation connections between encoder and decoder, as in U-Net, U-Net++, and CBAM-UNet, suffer from mismatched receptive fields where deep feature pixels correspond to shallow feature pixel areas, making it difficult for algorithms to balance feature fusion across different levels [?, ?], potentially leading to errors or ineffective liver region localization.

To address these challenges, this paper proposes MCAF-Net, which embeds Multi-feature Complementary Cross-Attention (MCCA) into the bottleneck layer. Through multi-feature fusion, MCCA achieves feature information complementation, generating rich and complete feature representations to reduce information loss. Cross-attention enables feature interaction to understand interdependencies between features, thereby ignoring interfering information such as noise. The encoder and decoder are connected by an Adaptive Multi-Scale Feature-fusion Module (AMFM), which, building upon MCCA' s reduction of noise and artifact effects, enables effective fusion across different layers. This allows the algorithm to fully acquire contextual and multi-scale information, alleviating over-segmentation while achieving fine segmentation of discontinuous liver regions and liver edges.

1.1 Medical Image Segmentation Tasks

The rapid development of artificial intelligence algorithms has made CNN-based medical image segmentation research increasingly prominent. Li et al. [**?**] proposed H-Dense-U-Net with coordinated connection network models. Albishri et al. [**?**] proposed a cascaded U-Net algorithm called CU-Net. Chen et al. [**?**] introduced TransUNet, which addresses U-Net' s limitations in modeling long-range dependencies and processing large images by incorporating a hybrid encoder based on the U-Net model. Cao et al. [**?**] proposed Swin-UNet, which follows the Swin Transformer architecture and computes self-attention within local windows to save computational complexity.

1.2 Attention Mechanisms

Attention mechanisms play a crucial role in medical image segmentation tasks. Channel and spatial attention mechanisms are widely used. Hu et al. [**?**] proposed Squeeze-and-Excitation Networks (SE-Net), which assign different channel weights to feature maps to learn the importance of different channels. Woo et al. [**?**] proposed the Convolutional Block Attention Module (CBAM), which employs both channel and spatial attention mechanisms to learn the importance of different channels and spaces. With the introduction of "Self-Attention," researchers began improving upon it. Gao et al. [**?**] proposed UTNet, which utilizes improved self-attention to capture long-range dependencies at different scales. Azad et al. [**?**] proposed DAE-Former, which achieves an alternative perspective through effective self-attention mechanism design.

2 Methodology

This section introduces the proposed algorithm for medical image segmentation tasks.

2.1 Overall Architecture of MCAF-Net

Medical images often contain noise from imaging equipment and artifacts caused by patient motion [**?**], which can affect algorithms and lead to erroneous segmentation of small, medium-sized, discontinuous, and blurred liver regions. Therefore, we propose MCAF-Net to achieve robust and accurate automatic liver segmentation in CT images. MCCA is embedded in the bottleneck layer, and the encoder and decoder are connected through AMFM to enable effective feature fusion. The overall structure of MCAF-Net is shown in [Figure 1: see original paper].

2.2 Adaptive Multi-Scale Feature-Fusion Module (AMFM)

As demonstrated in literature [**?**], feature fusion between encoder and decoder is an effective method for improving accuracy in high-level image tasks. However, simple fusion methods such as element-wise summation suffer from receptive field mismatch issues and struggle to balance contextual information across different levels through simple addition or concatenation operations. To alleviate this problem, we propose the Adaptive Multi-Scale Feature-fusion Module (AMFM), illustrated in [Figure 2: see original paper]. Input features undergo Spatial Pyramid Pooling Fusion (SPPF) [**?**], which generates multi-scale features through three cascaded 5$×$5 max pooling operations. The output features are weighted by the Channel Spatial and Multilayer perceptron Attention (CSMA) module. The weights w and $1 - w$ obtained from the CSMA module are combined with low-level and high-level features through weighted summation to achieve adaptive fusion. To deepen the network and reduce overfitting risk, inspired by ResNet' s residual connections, the initial input features of AMFM are added to the adaptively fused features, followed by a 1$×$1 convolution to output the final features. Compared with traditional simple feature fusion, AMFM effectively solves the feature mismatch problem between encoder and decoder, as deep feature pixels correspond to shallow feature pixel areas. The equivalent formulation of AMFM is given in Equation (1).

The architecture of the CSMA module is shown in [Figure 3: see original paper]. It primarily utilizes channel attention and spatial attention modules to obtain the importance of each channel and capture correlations between features of different dimensions, where X represents the input feature of the CSMA module. The computational formulas for the spatial attention module X_s and channel attention module X_c are given in Equations (2) and (3), respectively.

2.3 Multi-Feature Complementary Cross-Attention Mechanism (MCCA)

Since CT images often contain various noises and artifacts that can affect neural networks, we propose MCCA to identify and ignore noise and artifacts, thereby reducing their impact on the algorithm. The MCCA architecture is shown in Figure 4: see original paper. First, to reduce network complexity and overfitting

risk, input features pass through four different downsampling modules: D_g, D_s, D_m, and D_a. D_g performs grouped convolution downsampling to fuse local features and prevent the model from memorizing precise pixel locations, thereby improving generalization. D_s performs slice convolution downsampling to preserve initial data information. D_m and D_a perform max pooling and average pooling downsampling, respectively, to prevent loss of critical features. The obtained downsampled features are fused to generate rich and complete feature representations. Through multi-feature complementary fusion, when certain features are disturbed by noise or artifacts, the model can still make accurate judgments based on other features.

Currently, Self-Attention (SA) is a commonly used attention mechanism, employed as core components in both Trans-UNet and Swin-UNet. However, SA establishes global context models by computing relationships between each element position and all other positions. When input data contains noise, SA cannot effectively distinguish noise from useful features and may incorrectly incorporate noise into contextual information. Therefore, we introduce feature interaction to understand relationships and dependencies between different features. To achieve interaction between different features, we propose establishing cross-attention between two output features D_{mg} and D_{as}, as shown in Equations (5) and (6). During feature interaction, the model can automatically ignore noisy or redundant features and focus on important features by adjusting feature weights, thereby ignoring noise or artifacts.

The cross-attention mechanism can be described as mapping a set of key vectors and query-value vectors from another sequence to an output. The multi-head attention mechanism aims to capture relationships between different positions in the input sequence through multiple parallel attention heads and integrate this information. Each attention head performs cross-attention computation on input Query, Key, and Value, which are then concatenated to form the final output. The cross-attention mechanism is derived from Equations (7) and (8). The output is computed as a weighted sum of Values, where weights assigned to each Value are calculated from the scaled inner product of Query and corresponding Key. Specifically, by packing N Queries into $Q \in \mathbb{R}^{N\times d_k}$, each must first be matched with N Keys (K) using inner products. The dot product results between Query and Key are divided by $\sqrt{d_k}$ and normalized through a softmax function to obtain N attention weights. The final output is the weighted sum of N Values (V), where each $V \in \mathbb{R}^{N\times d_v}$.

The two output features from cross-attention undergo 1$×$1 convolution, with D'_{as} and D'_{mg} obtained through Equations (9) and (10), followed by fusion and upsampling to output the final MCCA module feature D', whose equivalent formula is given in Equation (11).

The SliceConv module is shown in Figure 4: see original paper. It splits adjacent pixels of input feature D to obtain four sub-features $D_{s1}, D_{s2}, D_{s3}, D_{s4}$, each half the size of the original matrix. The obtained feature maps are summed and passed through 1$×$1 convolution to output the final feature SliceConv, as

shown in Equation (12):

$$\text{SliceConv} = \text{BN}\left(C_{1\times1}\left(\sum_{i=1}^{4} D_{si}\right)\right)$$

where BN denotes Batch Normalization.

2.4 Loss Function

In convolutional neural networks, loss functions guide the network to minimize error. The Binary Cross-Entropy Loss (BCELoss) [**?**] primarily focuses on the probability of correct pixel classification, with the formula given in Equation (13):

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}[s_i \ln(b_i) + (1-s_i)\ln(1-b_i)]$$

where s_i and b_i represent the predicted value and ground truth, respectively.

3 Experiments and Results

This section presents systematic experiments and analyses to validate the effectiveness of the proposed modules and provide valuable insights for future research.

3.1 Datasets

shows the datasets and their partitions used in this study. We selected the Liver Tumor Segmentation dataset LiTS2017 (hereinafter "LiTS2017 dataset") [**?**] for liver and tumor segmentation, and used the 3Dircadb dataset to validate MCAF-Net' s generalization capability, both widely applied in liver surgery simulation and planning. Additionally, we utilized a real clinical dataset authorized by the Mayo Clinic, previously used in the "2016 NIH AAPM Mayo Clinic Low-Dose CT Grand Challenge" (hereinafter "LDCT dataset") [**?**]. This authoritative dataset has liver regions accurately delineated under the guidance of professional physicians from a top-tier municipal hospital, ensuring scientific validity and accuracy. Segmentation on the LDCT dataset was performed after denoising operations.

3.2 Experimental Environment

To ensure fairness, all experiments used identical settings: (1) NVIDIA GTX3090 with 24GB VRAM; (2) Adam optimizer; (3) learning rate of $1e^{-4}$; (4) 100 epochs.

3.3 Evaluation Metrics

This study employs Jaccard index, DSC, Recall, and Precision for semantic segmentation evaluation. TP represents true positives (correctly predicted positive samples), TN represents true negatives (correctly predicted negative samples), FN represents false negatives (positive samples predicted as negative), and FP represents false positives (negative samples predicted as positive).

Jaccard Index: Computes the ratio of intersection over union between predictions and ground truth. Values closer to 1 indicate better performance, as shown in Equation (14):

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

DSC (Dice Similarity Coefficient): Evaluates similarity between two sets. Values closer to 1 indicate better performance, as shown in Equation (15):

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

Recall: Measures the model' s ability to detect positive samples, as shown in Equation (16):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision: Measures model accuracy, as shown in Equation (17):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3.4 Comparison with Mainstream Networks

Numerous segmentation algorithms have been proposed for CT image liver segmentation. To further investigate MCAF-Net' s performance, we conduct experiments on LiTS2017 and 3Dircadb datasets, comparing quantitative and qualitative results with current mainstream algorithms.

[Figure 5: see original paper] shows segmentation results on four randomly selected CT images from LiTS2017, where boxes indicate regions requiring special attention and the bottom-right corners show magnified views. Quantitative analysis and visual results reveal that U-Net++ performs poorly, primarily because its simple feature fusion across different layers cannot capture complex dependencies between features, making it sensitive to regions with similar contrast. Both Trans-UNet and Swin-UNet exhibit misclassification, mainly

because Self-Attention simulates global dependencies through positional correlations, which may produce incorrect contextual dependencies when images contain noise or artifacts. Although SegFormer and UC-TransNet use Transformer as core components and multi-scale fusion strategies to compensate for incorrect contextual dependencies to some extent, their multi-scale fusion strategies often cause information loss and insufficient precision during pixel-level segmentation. nn-UNet employs multi-scale feature fusion technology, but its features become "diluted" or lost during multiple downsampling operations, ultimately causing information loss. Comparison of MCAF-Net with other mainstream algorithms in rows 1, 2, and 4 of [Figure 5: see original paper] shows that other methods missegment discontinuous liver regions due to artifacts causing blurred boundaries between tissues and liver edges, while MCAF-Net outperforms them. This is primarily attributed to MCCA reducing artifact and noise effects, while AMFM obtains more comprehensive contextual and multi-scale information. Row 3 demonstrates that other mainstream algorithms are more susceptible to interference, misclassifying other tissues as liver tissue, whereas MCAF-Net significantly reduces interference region effects due to MCCA incorporation. This further proves the effectiveness of multi-feature complementation and cross-attention in understanding feature interdependencies, reducing noise and artifact influence, and minimizing information loss.

Data in show that MCAF-Net achieves DSC and Jaccard of 96.24% and 92.83%, respectively, surpassing other mainstream networks. Both quantitative and qualitative results demonstrate MCAF-Net' s superior segmentation performance.

To evaluate MCAF-Net' s generalization capability, we directly apply models trained on LiTS2017 to the 3Dircadb dataset. shows MCAF-Net achieves DSC and Jaccard of 96.59% and 93.47%, respectively, outperforming other algorithms. [Figure 6: see original paper] presents segmentation results on randomly selected CT images from 3Dircadb. Similar to LiTS2017, other algorithms in rows 1 and 2 produce missegmentation due to artifact-blurred liver edges, while row 3 shows other methods misclassifying other tissues as liver. MCAF-Net effectively improves these phenomena, proving its ability to maintain noise/artifact suppression and fine liver edge segmentation across different datasets, thus demonstrating good generalization performance.

We further investigate MCAF-Net' s tumor segmentation performance on the LiTS2017 liver tumor dataset. Results are shown in [Figure 7: see original paper] and . Liver tumors typically exhibit diverse morphologies, blurred boundaries, and low contrast, increasing difficulty in distinguishing them from normal liver tissue. Noise further reduces this contrast, causing other algorithms to misclassify liver tissue as tumor tissue, particularly when detecting small lesions (e.g., Trans-UNet and Swin-UNet in row 1). This occurs because although they capture long-range dependencies, they cannot eliminate noise effects, reducing their ability to capture fine local tumor structures. U-Net++ exhibits numerous false detections due to simple feature fusion. SegFormer and UC-TransNet utilize

multi-scale fusion to eliminate noise effects and capture global information but fail to preserve detail information, especially for fine structures or edges.

3.5 Segmentation of Noisy CT Images

Low-Dose CT (LDCT) images contain significant speckle noise after radiation dose reduction, which further decreases contrast and blurs boundaries between liver and other tissues, increasing segmentation difficulty. Therefore, we investigate MCAF-Net's segmentation performance in noisy environments using LDCT images. Literature [**?**] demonstrates that U-shaped networks also perform well in image denoising. Consequently, we utilize MCAF-Net for LDCT denoising before segmentation. For the denoising task, we select Peak Signal-to-Noise Ratio (PSNR), Root Mean Square Error (RMSE), and Standard Deviation (SD) as evaluation metrics [**?**]. To quantitatively assess MCAF-Net' s denoising effectiveness, we measure RMSE, PSNR, and SD of SU-Net and Red-CNN on the dataset. [Figure 8: see original paper] shows PSNR and RMSE curves on LDCT, while presents final denoising results. Both curves and data demonstrate MCAF-Net' s superiority in LDCT image denoising compared to other algorithms.

Furthermore, we conduct quantitative comparisons with mainstream algorithms on the LDCT dataset, with results shown in . When segmenting denoised images, MCAF-Net' s DSC and Jaccard improve by 0.31% and 0.11% compared to non-denoised images, confirming that denoising benefits MCAF-Net' s segmentation performance on LDCT. [Figure 9: see original paper] shows final segmentation results of major algorithms on LDCT. Noise in LDCT images further reduces inter-tissue contrast, making U-Net++' s simple layer fusion more sensitive to similar contrast regions. Trans-UNet and Swin-UNet produce more incorrect contextual dependencies. SegFormer and UC-TransNet compensate for incorrect dependencies through multi-scale fusion but still misclassify liver tissue in this complex noisy environment. nn-UNet shows some anti-noise performance but remains affected by noise, causing over-segmentation. In rows 1, 3, and 4 of [Figure 9: see original paper], compared to LiTS2017, other mainstream algorithms are more vulnerable to noise and interference, misclassifying them as liver tissue, while MCAF-Net significantly reduces noise effects. Rows 1 and 2 demonstrate that with AMFM and MCCA, MCAF-Net not only reduces noise interference but also achieves fine segmentation of discontinuous liver tissue and edges.

Finally, box plots of all metrics on LDCT are shown in [Figure 10: see original paper]. MCAF-Net demonstrates significantly improved performance with smaller variance across all metrics. These results show that MCAF-Net maintains good segmentation performance even with substantial interference, proving its robustness and anti-noise capabilities.

4 Ablation Studies

To investigate the effects and contributions of key MCAF-Net modules, we conduct comprehensive ablation experiments on LiTS2017 and LDCT datasets. AFF-Net denotes the model containing only AMFM, while MFF-Net denotes the model containing only MCCA.

Ablation results are shown in [Figure 11: see original paper]. Comparison between AFF-Net and U-Net shows AFF-Net' s segmentation improves upon U-Net but remains suboptimal, primarily because enhanced contextual information perception only partially eliminates noise and artifact effects. Comparison among MFF-Net, AFF-Net, and U-Net demonstrates that adding MCCA significantly reduces interference from noisy or artifact-affected regions in both U-Net and AFF-Net, proving MCCA' s effectiveness in reducing information loss through feature complementation and promoting feature interaction via cross-attention, enabling the algorithm to ignore and identify noise or artifacts. Comparison between MCAF-Net and MFF-Net shows that based on MCCA' s effective reduction of interference, AMFM' s benefits become evident. Through adaptive feature fusion, AMFM effectively addresses encoder-decoder feature mismatch, enabling sufficient acquisition of contextual and multi-scale information for more accurate liver boundary identification.

Ablation results on LDCT are shown in , where MCAF-Net achieves DSC and Jaccard of 94.90% and 90.63%, representing improvements of 10.13% and 12.96% over U-Net, respectively.

To further explain module effectiveness, we visualize attention using heatmaps in [Figure 14: see original paper], where contours represent ground truth liver boundaries. Both U-Net and U-Net++ fail to effectively guide the model to focus on key regions due to simple additive feature fusion. CBAM-UNet adds a CBAM module to additive fusion, using a single weight w to indicate focus regions, but is less precise than AFF-Net (e.g., for discontinuous liver regions). AMFM generates w and $1-w$ guided by input feature content, and CSMA learns more accurate spatial weights, achieving effective feature fusion and guiding the model to accurately localize liver regions.

We further investigate the impact of cross-attention and multi-feature complementation on segmentation performance, with results shown in [Figure 12: see original paper] and . S-UNet represents the bottleneck embedding Self-Attention module, while C-UNet represents the bottleneck embedding Cross-Attention module. Results show C-UNet significantly reduces interference from noisy or artifact-affected regions compared to S-UNet, demonstrating cross-attention' s superior anti-noise capability on this dataset. MCAF-Net achieves more precise liver localization than C-UNet and further reduces noise/artifact effects, proving that feature fusion generates rich, complete feature representations that reduce information loss. tests segmentation accuracy with different complementary modules: MCCA-G (without Gconv), MCCA-A (without average pooling), MCCA-M (without max pooling), and MCCA-S (without SliceConv). Results

show 0.1%-0.5% accuracy drop when any module is excluded, while inclusion of all four modules yields significantly higher accuracy, confirming the necessity of all four modules in multi-feature complementation.

Additionally, we study the impact of Multi-Head Cross-Attention (MHCA) head count and MCCA module quantity on MCAF-Net performance on LiTS2017. We test four head counts (2, 4, 8, 16) and different MCCA quantities (1, 2, 3, 4) embedded in the bottleneck. Results in [Figure 13: see original paper] show DSC decreases with 16 heads, primarily because while more heads enhance interaction between sequences, excessive head count causes model overfitting. Similarly, DSC peaks when 3 MCCA modules are embedded, decreasing with 4 modules. Although more MCCA modules reduce noise effects and enhance sequence interaction, excessive modules also cause overfitting. Therefore, MCAF-Net achieves optimal performance for liver segmentation with 8 heads and 3 MCCA modules.

Conclusion

Medical images are affected by equipment and patient factors that generate noise and artifacts, and neural network algorithms are susceptible to such interference. Accurate and robust automatic segmentation of liver regions in CT images is an urgent problem with significant clinical value. To address this, we propose MCAF-Net.

On LiTS2017, MCAF-Net achieves DSC and Jaccard of 96.24% and 92.83%, respectively, demonstrating superior performance compared to other mainstream algorithms. Ablation experiments reveal that performance improvements primarily benefit from: (1) MCCA reducing information loss through multi-feature complementation and dynamically adjusting feature weights via cross-attention to suppress noise and artifact features; and (2) AMFM enabling effective feature fusion through SPPF and adaptive encoder-decoder feature fusion, allowing the network to acquire rich contextual and multi-scale information for improved liver boundary perception. When combined, AMFM more accurately identifies liver tissue boundaries while reducing interference from non-liver tissues. On 3Dircadb, DSC and Jaccard reach 96.59% and 93.47%, respectively, maintaining superiority in ignoring artifacts/noise and identifying liver edges, thus proving MCAF-Net' s generalization capability.

We also evaluate MCAF-Net' s performance in noisy environments using the LDCT dataset. With professional physician guidance, we accurately and scientifically delineated liver tissue labels on LDCT abdominal images and proposed a denoise-then-segment approach. Results show DSC and Jaccard of 94.90% and 90.63% on LDCT, improving over U-Net by 10.13% and 12.96%, respectively. Compared to other mainstream networks, MCAF-Net maintains excellent segmentation performance in noisy environments and demonstrates good denoising capability with PSNR reaching 32.61 (0.77 higher than SU-Net), proving its robustness and anti-noise performance.

However, MCAF-Net has room for improvement. shows parameter count and

inference time comparisons: MCAF-Net offers favorable inference time but does not have an advantage in parameter count. Future research will focus on designing lightweight network models. Additionally, while MCAF-Net excels at obtaining outer liver boundary information, it does not treat vessels and other non-liver tissues inside the liver as liver tissue. However, since some internal tissues adhere closely to liver tissue, fine segmentation remains challenging, as shown in [Figure 15: see original paper]. Future work will investigate effective separation of internal non-liver tissues at the algorithmic level.

In summary, we propose MCAF-Net for accurate and robust liver CT image segmentation. By embedding MCCA in the bottleneck layer, the network generates rich, complete feature representations through multi-feature fusion, reduces information loss, and employs cross-attention for feature interaction to dynamically adjust weights and suppress noise/artifact features. AMFM enables fine liver edge segmentation by adaptively fusing encoder-decoder features through SPPF, allowing the algorithm to obtain rich contextual and multi-scale information. On LiTS2017, DSC and Jaccard reach 96.24% and 92.83%; on 3Dircadb, they reach 96.59% and 93.47%. Compared to mainstream algorithms, MCAF-Net shows superior segmentation performance with better accuracy for blurred edges, similar contrast regions, and small, discontinuous liver areas. Ablation experiments on LiTS2017 confirm AMFM and MCCA effectiveness in detail extraction and anti-interference. Finally, experiments on noisy LDCT images demonstrate that MCAF-Net maintains strong segmentation performance in noisy environments and shows good denoising capability, proving its robustness and anti-noise properties.

Author Contributions: Shen Tong collected and organized data and drafted and revised the manuscript; Wang Jianlin designed the overall program code; Gu Runqiu proposed and designed the research; Wang Yadi collected literature.

References: (Preserved exactly as in original)

Note: Figure translations are in progress. See original paper for figures.