SPA-DCN-NET: A Gated Multi-Scale Local Contrast Normalization Network for Ultrasound Image Segmentation of Liver

Su Ming Jian, Afizan Bin Azman
School of Computer Science, Taylor's University Lakeside Campus, Subang Jaya, Malaysia

***Abstract*—Segmentation of the liver in ultrasound images is a critical task in medical image analysis, yet it remains challenging due to acoustic speckle noise, brightness instability, and deformations caused by probe pressure. To address these problems, this study presents SPA-DCN-NET, a lightweight framework that integrates three synergized components: first, a learnable gated local contrast normalization (Gated LCN) module utilizes a sigmoid soft-gate mechanism to dynamically fuse LCN-enhanced features with original features, effectively stabilizing the features for training. Second, a spatial pyramid attention (SPA) module applies multi-scale context aggregation, transforming the clear features provided by Gated LCN into deformable convolutional networks. Third, these features guide deformable convolutional networks to adaptively adjust sampling grids, ensuring precise delineation of irregular liver boundaries in ultrasound images. Experimental results demonstrate that our SPA-DCN-NET achieved mean IoU scores of 83.52%, 75.57%, 73.85%, and 85.94% across the four datasets, respectively. These results are all higher than those obtained by UNet, nnUNet, and ResUNet. The metrics indicate that our SPA-DCN-NET is more adaptable to the ultrasonic medical environment compared to other existing medical segmentation networks, and it is a recommended network of image analysis for ultrasound abdominal scans.**

***Keywords*—Spatial pyramid attention; deformable convolutional network; ultrasound image; medical image analysis; soft-gate mechanism; local contrast normalization**

I. Introduction

Liver disease is a major burden on human health. In 2019, the global mortality attributed to hepatitis B and hepatitis C surpassed one million [1]. Liver diseases often present a diagnostic challenge due to the inherent limitations of current detection modalities. Ultrasonography remains the predominant imaging modality for hepatic assessment. However, during the scan, the probe will repeatedly press against the skin. These actions result in dynamic alterations to both hepatic morphology and the intensity of parenchymal echoes, concomitantly causing distortion of the ultrasonic beam's propagation path, and ultimately lead to instability in the image's structural representation and brightness. Research has demonstrated that inappropriate calibration of time gain compensation (TGC) settings can dramatically impair AI model performance, manifesting as a substantial decline in diagnostic accuracy, ranging from 5% to 40% [2].

At present, there are already some medical image segmentation networks, such as UNet [3] and nnUNet [4], whose core module is a convolutional neural network (CNN). Multiple studies have shown that the classification, detection, and segmentation performance of CNNs will significantly decline under different lighting conditions. For instance, pre-trained networks such as VGG16 and ResNet exhibit significant instability under 15 different lighting conditions, indicating that existing CNN models lack robustness to lighting changes and need to enhance their adaptability through means such as data augmentation [5][6][7]. The fundamental reason for CNN's sensitivity to lighting changes is its reliance on low-level features of the input image, such as brightness, contrast, and texture. Research has found that CNN is extremely sensitive to minor changes such as high-frequency disturbances and non-uniform illumination, and is easily misled by non-uniform illumination, resulting in a significant drop in classification accuracy. The introduction of a biological-inspired anterior cortex module can significantly enhance the robustness of CNN to global light intensity and contrast variations [8]. Introducing diverse lighting conditions in the training set or synthesizing weak light or non-uniform lighting images can effectively enhance the model's adaptability to actual lighting changes [9]

The local contrast normalization (LCN) is a classic feature enhancement method, and its design inspiration comes from the neural mechanism of the human visual system, specifically the concept of lateral inhibition. In the retina, excited neurons inhibit neighboring neurons, which naturally enhances the contrast between features and their surrounding background, thereby improving edge detection and pattern recognition. LCN mathematically simulates this process by transforming the pixel intensity values based on the properties of their local neighborhood [10]. LCN first uses local mean subtraction to remove average local brightness and center the image data around zero. Second, the resulting data is normalized by the local standard deviation. By normalizing intensity based on local statistics, LCN effectively standardizes the image contrast across different regions. This stabilization is crucial in medical imaging, particularly for mitigating the aforementioned brightness and contrast instability problems common in ultrasonography. LCN helps deep learning models become more robust to non-uniform illumination and poor contrast conditions, which is essential for maintaining high diagnostic accuracy in a diverse clinical setting. This makes it a valuable pre-processing step before inputting images into a computer-aided diagnosis system.

Normally, the intensity of echo propagation during medical examination can be expressed as Eq. (1):

$$I_{echo}(P) \approx I_{in} \cdot R^{'}(P) \cdot e^{-2\alpha \cdot f(d - \Delta d(P))} \tag{1}$$

This equation states that the final image brightness I_{echo} is the result of the initial energy I_{in} being reflected by the tissue $R^{'}$ and then exponentially attenuated by the distance it traveled d, the wave frequency f, and the tissue's absorption properties α. The exponential term e^{*} in Eq. (1) indicates that signal intensity inherently drops with depth and tissue density. This physical limitation causes the brightness instability that standard CNNs fail to handle. To specifically counteract this non-linear attenuation described in Eq. (1), we introduce the gated local contrast normalization (Gated LCN) module to dynamically re-calibrate the feature intensity.

Some of the currently popular medical image segmentation frameworks, such as UNet, struggle to track objects with complex or soft regions in ultrasound images due to the limitations of the receptive field of the convolutional neural network [11]. To tackle this problem, deformable convolutional networks (DCN) [12] lie in that the sampling points of its convolutional kernels are learnable and can adaptively adjust the sampling position and shape according to the input features. Previous literature [13] has demonstrated that integrating DCN into image segmentation networks can improve the effect of organ segmentation. However, none of their studies focused on the changes in the brightness of ultrasound images. Our SPA-DCN-NET resolves the challenge of chaotic brightness instability and non-rigid geometric deformation in a unified, learnable framework. SPA-DCN-NET introduces a cooperative adaptive mechanism, uses Gated LCN as a dynamic valve to adjust the contrast, and guides the spatial pyramid attention-driven deformable convolution to adjust the sampling mesh. This end-to-end loop realizes the joint optimization of light steady-state and geometric accuracy, thereby achieving more precise segmentation under complex artifacts.

In this study, the spatial pyramid was applied to address the CNN model's reliance on a fixed input size while enhancing the model's scale invariance. In [14] the spatial pyramid attention (SPA) module can be directly placed in the backbone network and enhances the performance of the entire network. During the research, we found that directly placing LCN in front of SPA, although it can alleviate the problem of unstable image brightness, may overly suppress the information crucial for segmentation in the original features, leading to a decline in the algorithm's performance on normal datasets. In order to solve this problem, we introduce a learnable gated mechanism to regulate the influence of LCN features on the SPA module.

MobileNetV3 [15] is an efficient, lightweight network that is smaller than EfficientNet [16]. Since our decoder includes SPA, Gated LCN, and DCN, using MobileNet can save resources for the overall operation of the framework. Furthermore, we have simple concatenation between symmetrical stages of the encoder/decoder, a structure that enables effective pixel-based training in limited medical data [17]. Inspired by the work of [18], we proposed a model specifically designed for abdominal ultrasound scanning called SPA-DCN-NET, which integrates deformable convolutional networks (DCN) and integrates the gate control brightness function for SPA modules. Spatial pyramid pooling (SPP) overcomes the limitation of fixed input image size in convolutional neural networks, realizes multi-scale feature extraction, and improves computational efficiency [14]. Attention mechanism dynamically focuses on spatial and feature channels that are more important for the current task [19].

The study proposes SPA-DCN-NET, a lightweight and efficient end-to-end architecture that includes MobileNet and the GatedLCN-SPA-DCN module, specifically designed to address the issues of unstable brightness and low contrast. Our core innovations and contributions are structured as follows:

Proposal of learnable gated local contrast normalization (Gated LCN). To tackle the unstable brightness and contrast in ultrasound images, we designed a novel gating mechanism based on sigmoid soft gates. This mechanism dynamically fuses the contrast-enhanced features from local contrast normalization with the original features, effectively stabilizing image quality and yielding high-quality feature maps.

1) Improvement of the spatial pyramid attention module. To avoid the loss of spatial details typically caused by global average pooling in conventional attention mechanisms, we construct the improved SPA module using multi-scale dilated convolution. This preserves high-resolution contextual information. Combined with the refined features output by Gated LCN, the attention can be more accurately focused on low-contrast areas, such as the liver boundary, significantly enhancing segmentation quality.

2) Construction of the lightweight and efficient MobileNet + GatedLCN-SPA-DCN architecture. We adopt MobileNetV3 as the encoder and deeply integrate the joint GatedLCN-SPA-DCN module into the decoder. This architecture achieves a superior balance between efficiency and accuracy, significantly outperforming mainstream segmentation models, which makes it robust, accurate, and highly suitable for clinical deployment.

In summary, the SPA-DCN-NET achieves innovation not only at the module level but also provides an efficient, robust, and accurate end-to-end system-level solution for medical ultrasound image segmentation.

II. Related Works

A. Encoder-Decoder Structure and Skip Connection

Our SPA-DCN-NET is based on the encoder-decoder architecture. This structure is designed to effectively capture the contextual information of the image while precisely locating the target boundaries. The encoder, similar to a seq-to-seq function [20], compresses variable input data into a dense feature sequence. The decoder then performs upsampling operations, typically doubling the feature map size at each layer until the original image resolution is reached. Finally, a

1*1 convolution kernel adjusts the channel number to the required number of categories, followed by the softmax function to generate the final probability map. In the classic UNet structure, skip connections are employed between corresponding layers of the encoder and decoder. While some UNet-like architectures, such as ResUNet [21] and nnUNet, have achieved success by deepening networks or optimizing heuristics, they fundamentally operate as static, linear pipelines. They process images with stable brightness parameters. However, in a clinical ultrasound setting, artifacts such as acoustic shadowing and TGC variation cause dynamic brightness instability [2]. Existing encoder-decoder frameworks lack a dynamic adaptation mechanism to recalibrate feature intensity on-the-fly, often leading to boundary prediction errors in low-contrast regions. We have effectively solved this problem by using a dynamic gating mechanism. This mechanism allows the decoder to collect as much fine-grained spatial information as possible.

B. Lightweight and Efficient Backbone Network

The encoder-decoder architecture offers flexibility in selecting the backbone network for the encoder position. Options include efficient and lightweight networks like MobileNet, ShuffleNet, EfficientNet, and GhostNet, or resource-dependent networks such as ResNet, VGG, and DenseNet. We utilize MobileNetV3 due to its efficiency, and from its use of 1*1 convolutions, which are more resource-friendly than UNet's typical 3*3 convolutions. MobileNetV3 inherently supports multi-layer feature extraction, providing feature maps of different resolutions that can naturally serve as the input for the skip connection at each level of the decoder.

C. Spatial Pyramid Attention Network

The spatial pyramid block is crucial in deep learning for handling multi-scale information and addressing the challenge of unfixed input size. Spatial pyramid pooling (SPP) [14] changes the receptive field by altering the size of the pooling grid. Atrous spatial pyramid pooling (ASPP) [22] achieves this by modifying the dilation rate. Traditional attention mechanisms, such as Squeeze-and-Excitation networks (SENet) [19], typically rely on global average pooling (GAP) to compress 3D feature maps into 1D channel vectors to obtain channel dependencies. However, this method risks losing important spatial structure information. The spatial pyramid attention module is specifically designed to overcome this limitation by combining structural information retention and structural regularization. When calculating attention weights, it not only focuses on feature channels but also utilizes the spatial structure information within the feature map. Related papers include the SPANet, the SPAN [23], the SAPNet [24], and the MSPANet [25]. While SPANET and MSPANET integrate channel attention, and SPAN and SAPNET integrate spatial self-attention, the attention formulation in all four shares the common starting point of applying the traditional GAP. GAP compresses the entire spatial dimension into a single value. Although this helps prevent overfitting, it seriously compromises spatial structure information, especially with large feature maps.

Gou et al. in [14] replace GAP with a spatial pyramid structure, integrating multi-scale feature extraction, structural information retention, and long-distance channel dependence modeling into a unified module. In [24], task-specific guiding information, such as the classification confidence graph or the segmentation prediction graph, is introduced to generate spatial attention masks. This mask dynamically assigns weights to the feature maps of each scale in the pyramid, allowing the network to focus on the regions and scales most discriminative for the current task. The self-attention blocks of each layer in [23] utilize different dilation rates. This enables each pixel at the top layer to perceive a very large area at the bottom layer. Furthermore, positional projection is introduced to better encode the relative spatial positions of neighboring pixels.

D. Local Contrast Normalization

The idea of local contrast normalization was proposed in [27]. Deep learning variants of LCN [26] aim to address the issue of insufficient robustness caused by fixed LCN parameters, making them suitable for solving illumination and gain adaptation problems in ultrasound images. In [28], a universal local context normalization layer was proposed, which can be inserted into the learnable modules of modern deep networks. In [26], subtraction-type LCN and division-type LCN were proposed, both capable of eliminating low-frequency illumination variations.

E. Gating Mechanism

LCN can enhance the model's robustness on datasets with significant light and shadow variations, but it may sacrifice accuracy on normal datasets. Gated LCN enables the network to adapt to switch between the robustness of LCN and the accuracy of the original feature. Attention gate and dilation U-shaped network (GDUNet) proposed by Chen et al. [29] is an outstanding representative of this idea, introducing lightweight attention gates (AG) in the UNet skip connections. The AG module effectively filters out low-level semantic noise in the spatial and channel dimensions, thereby attenuating irrelevant features. Zeng et al. proposed a deeply supervised attention-gated V-Net (DAG V-Net) for automatic fetal head circumference measurement [30]. This model integrates the attention gating module into the 2D V-Net architecture to optimize skip connections. Schlemper et al. applied the AG model to standard section detection in fetal ultrasound, constructing Ag-Sononet [31]. This model applies an attention gate to the deep feature map of Sononet (a VGG variant) to generate an attention map. These attention maps are used to guide the final classification decision via adaptive feature pooling and can be directly applied to weakly supervised target localization.

Although LCN effectively alleviated the unevenness of light exposure, overuse of LCN can also have significant side effects. It weakens the texture details in images, reducing segmentation accuracy in normal scenarios. Among the existing medical segmentation methods, once the network parameters are fixed, it is difficult to ensure that the data with uneven brightness will have a stable segmentation accuracy rate. In contrast, our proposed Gated LCN acts as a learnable valve. Unlike standard attention gates that focus on spatial regions, our gating mechanism generates the weights (0 to 1) using the sigmoid function. The network can then automatically learn to suppress background noise, ensuring the network only activates normalization when necessary.

F. Deformable Convolution

To compensate for the inherent limitations of ordinary convolution in the receptive field, Dai et al. proposed deformable convnets (DCNv1) [13]. By introducing learnable offsets in the standard convolution and ROI pooling modules, the network can adaptively adjust its sampling position according to the input content. To further explore the potential of deformable convolution, Zhu et al. proposed deformable convnets v2 (DCNv2) [32]. DCNv2 includes an additional modulation mechanism compared to DCNv1, based on the learnable offset. A learnable modulation scalar is introduced for each sampling point to dynamically adjust the characteristic amplitude of that sampling point. This enables the network to simultaneously consider the location and weight of the sampling points. Our approach couples the SPA with DCN. By feeding DCN with features that are already structurally emphasized by SPA, we ensure that the geometric deformation is guided by meaningful semantic context rather than noise.

G. Summary of Research Gap and Our Contribution

In summary, existing works predominantly address the challenges of segmentation through either deeper static architectures (e.g., ResUNet) or standalone modules (e.g., standard DCN or LCN). The critical limitation in the current literature is that this work did not integrate these functions together. Traditional linear pipelines fail to adapt when both situations occur simultaneously—a common scenario in handheld ultrasound due to probe pressure and gain shifts. SPA-DCN-NET does not merely apply these techniques in isolation but links them sequentially. The Gated LCN stabilizes brightness to allow the SPA to correctly focus attention, which in turn guides the DCN to deform accurately. This integrated design can better adapt to the actual medical setting of ultrasound segmentation.

III. Methodology

A. Functional Mechanisms of SPA-DCN-NET vs. Baselines

We selected three common medical image segmentation networks, namely UNet, nnUnet, and ResUNet. These models perform well on general medical images, but none of them can solve the problem of uneven illumination and simultaneous geometric deformation in ultrasound images. Unlike these benchmark models, our SPA-DCN-NET introduces a dynamic cooperative loop to achieve synchronous adaptive adjustment of brightness and shape. The functional differences between our SPA-DCN-NET and three other common medical image segmentation networks are summarized in Table I.

TABLE I. A Function Comparison of the Core Differences of SPA-DCN-NET and Other 3 Popular Medical Segmentation Networks

	Standard UNet	**ResUNet**	**nnUNet**	**SPA-DCN-NET (Ours)**
Process Flow	Linear. Input → Conv → Output	Linear. ResBlock → Deep Features	Linear. Adaptive Preprocessing → UNet	Synergistic. Gating LCN↔ Spatial Attention ↔ Deformable Shape
Illumination Adaptation	Static. Rely on fixed input preprocessing or data augmentation.	Static. Features are extracted relying on residual connections, without a specific lighting adaptation module.	Static. Fixed normalization based on the statistics of the entire dataset.	Dynamic Gating. Pixel-by-pixel trade-off between original features and normalized features, adaptively resisting local brightness changes.
Geometric Deformation	Fixed grid Standard 3×3 convolution with a fixed receptive field shape.	Fixed Grid. Although there are residual connections to deepen the network, the geometry of the convolutional kernels cannot be changed.	Fixed Grid. The standard UNet architecture relies on data augmentation to handle deformation.	Adaptive Shape. DCN v2 dynamically adjusts the Offset of the convolution kernel sampling points based on the features extracted by SPA to precisely fit the irregular liver boundary.
Feature Focus	Hard coding, Simple concatenation relying on Skip Connection.	Implicit. Features are implicitly learned through deep networks without explicit attention.	None. Regression to pure convolution.	Multi-scale Attention. The SPA module utilizes dilated convolution to capture multi-scale contexts and explicitly guides the DCN to focus on the lesion area.
Ultrasound Specificity	Weak. General model, sensitive to noise.	Medium. The residual structure alleviates vanishing gradients but does not solve artifacts.	Medium. Depending on the amount of data, there is no specific design for specific ultrasonic artifacts.	Strong. Specifically designed for ultrasonic Speckle noise and Probe Pressure deformation.

B. Network Overview

The SPA-DCN-NET framework is built upon the principles of the encoder-decoder architecture, utilizing progressive downsampling and upsampling for robust segmentation. We employ MobileNetV3-Small as the downsampling backbone, a choice driven by its high computational efficiency and low latency, making it highly suitable for real-time clinical ultrasound scenarios. The network's initial layer is modified to seamlessly accept a single-channel input. The inherent layer-by-layer downsampling of the MobileNetV3 backbone naturally generates multi-resolution feature maps at scales H/2, H/4, H/8, H/16, and H/32. These multi-level features are captured to establish skip connections with the decoder, a design that perfectly embodies the core mechanism of the UNet design. In the decoder section, upsampling operations are performed, and at each corresponding layer, the Gated LCN block, the SPA block, and DCN are cascaded and jointly construct the EnhancedDeformConv block. The complete structural flowchart of the proposed SPA-DCN-NET is illustrated in Fig. 1.

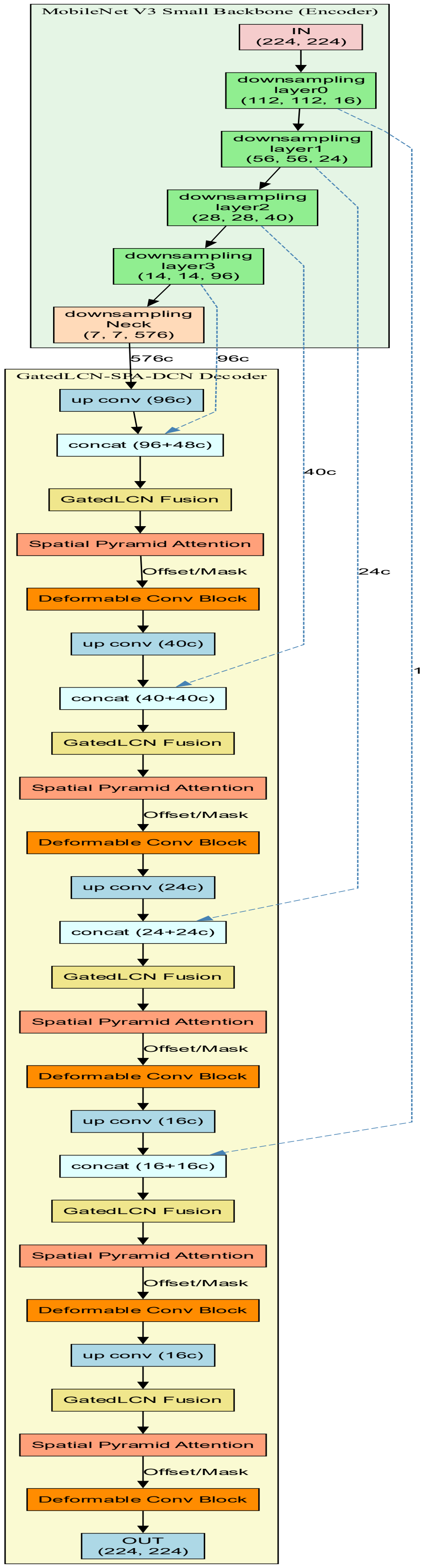

Fig. 1. The complete structural flowchart of the proposed SPA-DCN-NET.

C. Encoder and Decoder

To adapt to single-channel medical images, the weights of the first layer convolutional kernels are specially initialized. From the encoder's features module, we precisely captured the intermediate feature maps of five key layers for subsequent skip connections. These feature maps are downsampled step-by-step in space, and the number of channel counts is increased to 16, 24, 40, 48, 96, forming a pyramid-shaped feature representation from detail to semantics. The decoder consists of a total of five upsampling stages, gradually restoring the feature map from H/32 to H. In the final stage, since the highest resolution has been reached, no new skip connections are introduced. Only the EnhancedDeformConvBlock is used for the final refinement of the features. Ultimately, a 1x1 convolutional layer maps the number of channels of the refined feature map to the number of categories (such as background and liver), outputs the logits of each pixel belonging to each category, and completes the segmentation task. Through this encoder-decoder collaboration, skip connection fusion, and SPA-DCN adaptive refinement design, the model can achieve high-precision segmentation of complex deformed livers while maintaining high efficiency. The concatenated feature maps are sent to the EnhancedDeformConvBlock for processing. This is the core innovative module of this model, which seamlessly integrates spatial pyramid attention (SPA) with deformable convolution (DCN). First, the module applies gated local contrast normalization (Gated LCN Fusion) to the input features. This module adaptively performs dynamic weighted fusion between the illumination robustness of LCN-enhanced features and the accuracy of the original features through the sigmoid soft gate. This enables the model to resist the complex light and shadow changes commonly found in ultrasound images and stabilize the feature representation. The SPA module conducts multi-scale analysis on the input features to generate an attention weight map, highlighting the spatial regions that are more important for the liver segmentation task. The DCN module utilizes the features enhanced by SPA to predict offsets and modulation_masks, and then performs deformable convolution on the original concatenated features. This enables the sampling mesh of the convolutional kernel to adaptively adjust according to the shape of the liver, thereby more accurately capturing its irregular boundaries.

D. Soft Gated Local Contrast Normalization Layer

The mathematical form of local contrast normalization is:

$$LCN(x) = \frac{x - \mu_{local}}{\sigma_{local} + \varepsilon} \tag{2}$$

In Eq. (2), μ_{local} is the local mean, σ_{local} is the local standard deviation. The step of subtracting the local mean μ_{local} is equivalent to removing the local low-frequency component from the original signal. Retain high-frequency information such as edges, textures, and details that change rapidly, and suppress the low-frequency background in smooth areas. This is consistent with the effect of high-pass filtering. Therefore, LCN(x) mainly retains structures with strong local contrast, such as boundaries, spots, and textures, which belong to high-frequency components in the frequency domain. The original feature x contains complete spectral information, but it is mainly low-frequency. The schematic diagram of Gated LCN is shown in Fig. 2. We use the sigmoid soft gate to achieve dynamic scaling and modulation of LCN features through element-wise multiplication. There are two independent sigmoid gateways, one is based on X_LCN, the other one is based on X, and they form a pair of approximately complementary soft gates in a 1.0− σ manner, thereby allowing the network to perform adaptive, channel-by-channel or spatially position-by-space weighted fusion between the robustness of LCN and the detail accuracy of X. Here we have employed sigmoid soft gate technology. The sigmoid soft gate does not have a fully open (1.0) or fully closed (0.0) state. The sigmoid soft gate achieves dynamic scaling and modulation of LCN features through element-wise multiplication. The dynamic fusion process is mathematically formulated in Eq. (3):

$$Output = G_{LCN} \odot \mathrm{X}_{\mathrm{LCN}} + G_{identity} \odot \mathrm{X} \tag{3}$$

where, $\odot$ denotes element-wise multiplication, and the complementary gate is defined as $G_{identity} = 1.0 - Sigmoid(Conv_{1*1}(X))$.

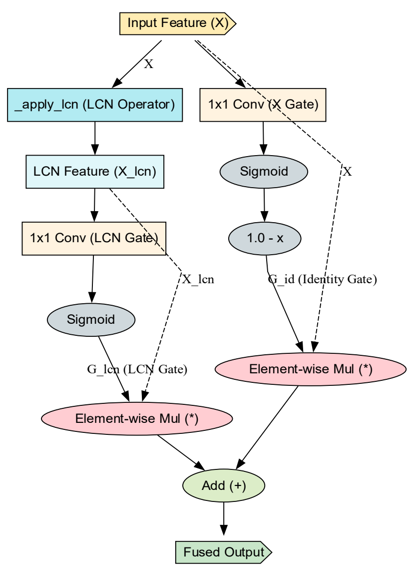

Fig. 2. Data flow in local contrast normalization block. The network adaptively learns the optimal fusion ratio between LCN-enhanced and original features at each spatial position, driven by local contrast and texture information.

As described in Eq. (3), this equation functions as a dynamic arbitration mechanism. When the network detects significant brightness instability or high-frequency noise (typical of acoustic attenuation), the learnable weight G_{LCN} tends towards 1. In this scenario, Eq. (3) prioritizes the normalized features X_{LCN} to ensure model robustness. Conversely, in regions with high signal-to-noise ratios, $G_{identity}$ allows the original feature stream X to pass through to preserve fine anatomical textures. Thus, Eq. (3) effectively prevents the over-normalization problem by selectively activating LCN only where necessary.

E. Spatial Pyramid Attention Block

Spatial pyramid attention was explicitly proposed in the papers [14], [22], [23], [24]. We mainly referred to [14], however, our SPA module did not use the pooling layer. Instead, multi-scale features are captured through dilated convolution layers with different dilation rates. As shown in Fig. 3, dilated convolution is used instead of adaptive average pooling because it expands the receptive field without losing resolution. It still performs convolution operations on each spatial position, thereby being able to capture fine spatial details. Although adaptive average pooling can extract multi-scale information, it is essentially a downsampling operation. During the pooling process, a large amount of spatial detail information is lost.

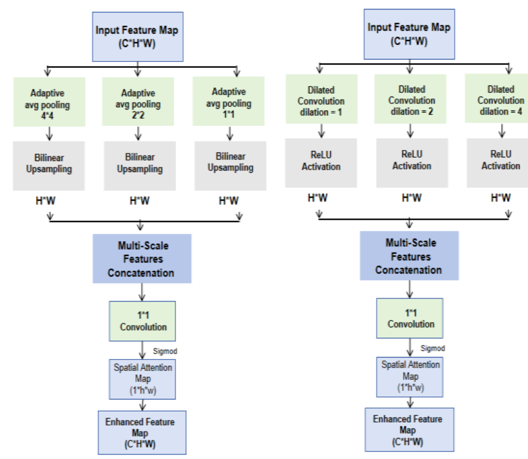

Fig. 3. The structure of our spatial pyramid attention. The pooling layer from [14] (left) are replaced by dilated convolution layer (right).

F. Modulation Convolution

Modulation convolution is mainly used to generate modulation masks, providing an additional intelligent weighting mechanism for deformable convolutional networks (DCN). The modulation mechanism assigns a learnable weight to each adaptive sampling point. It receives features enhanced by SPA as input, and a convolutional layer followed by a sigmoid function is applied to compress the output value to the range of [0, 1], forming a modulation mask. When DeformConv2d is executed, it multiplies the sampled features element by element by the corresponding modulation weights. The feature contribution of this sampling point is high when the weight is close to 1. When the weight approaches 0, the feature contribution of this sampling point is suppressed. Modulation convolution enables DCN not only to adaptively select the sampling position, but also to adaptively determine the importance of each sampling point.

G. Deformable Convolution Implementation

The deformable convolution (DCN) is strategically placed in the decoder's upsampling path, where it takes input from the combined encoder skip connection and the decoder's upsampled feature map. Its core role is to employ adaptive sampling to precisely capture and restore fine target shape details as the feature map resolution grows, thereby enhancing final segmentation accuracy. This adaptability relies on a specialized convolutional layer that learns the necessary spatial shifts for every sampling point throughout the training process. In the forward propagation, these predicted shifts are used in the DCN calculation. The resulting prediction is measured against the ground truth mask to quantify the loss. During back propagation, the loss signal propagates backward, updating all layer weights. If a specific shift leads to reduced loss, the weights responsible for generating that shift are reinforced. Through iterative training, the prediction layer learns precise spatial adjustments to features, which minimizes segmentation error and accurately refines object boundaries (see Fig. 4).

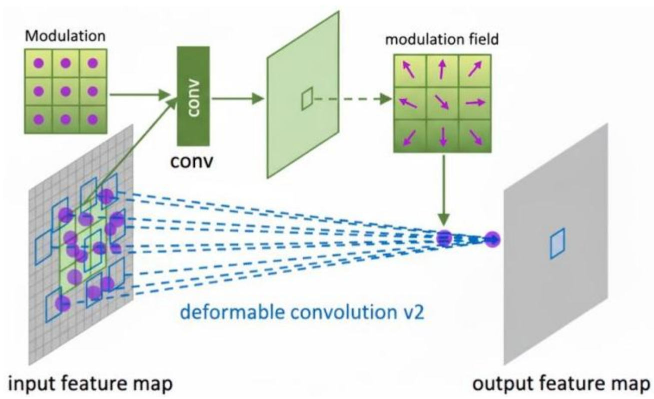

Fig. 4. The working mechanism of deformable convolution v2. By learning offset and modulation fields from the input, it enables the kernel to achieve position and amplitude adaptation. The offsets (blue box) adjust sampling points to focus on target boundaries, while modulation values determine feature importance for dynamic weighting. This allows DCNv2 to precisely capture the shape and details of complex targets.

H. The Combination of Spatial Pyramid Attention and Deformable Convolution Module

The mathematical expression of the standard convolution is:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{4}$$

In Eq. (4), R is the sampling mesh of the convolution kernel, w is the weight, and p_0 is the position on the output feature map.

Deformable convolution, based on standard convolution, adds a learnable offset Δp_n and a modulation scalar m_n to each sampling point p_n. Since $p_0 + p_n + \Delta p_n$ is usually a fractional coordinate, it is calculated through bilinear interpolation:

$$y_{DCN}(p_0) = \sum_{p_n \in R} w(p_n) \cdot x_{SPA}(p_0 + p_n + \Delta p_n) \cdot m_n \tag{5}$$

where, Eq. (5) is the output of deformable convolution. $p = p_0 + p_n + \Delta p_n$ is the fraction position of grids.

$$x_{SPA}(p) = \sum_{q} G(q, p) \cdot x_{SPA}(q) \tag{6}$$

In Eq. (6), q represents all integer positions on the feature map, while $G(\cdot,\cdot)$ is the bilinear interpolation kernel. The final EnhancedDeformConvBlock output y in Eq. (7) is obtained through deformable convolution, batch normalization, and the ReLU activation function:

$$y = relu(batch\_norm(y_{DCN})) \tag{7}$$

Fig. 5 illustrates the data flow in the whole combination of spatial pyramid attention and deformable convolution module.

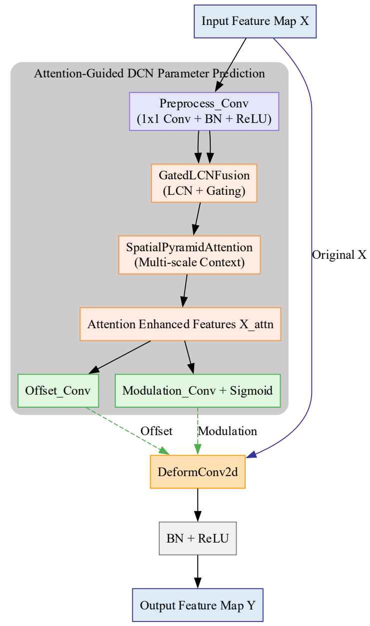

Fig. 5. Schematic diagram of the combination of spatial pyramid attention and deformable convolution module.

I. Loss Function and Optimizer

To effectively train the MobileNet + GatedLCN-SPA-DCN architecture, we utilize the Cross Entropy Loss function paired with the Adam optimizer. The loss function is chosen specifically to combat the class imbalance problem in medical segmentation, where background pixels vastly outnumber target pixels. By supplying a weight vector class_weights_tensor as a parameter, the function applies a higher penalty to misclassified minority pixels, thus forcing the network to focus intently on accurate boundary delineation and shape identification of the target. This weighted approach is critical for enhancing segmentation accuracy on the target region and improving metrics like the Dice Score. The Adam optimizer [33] is employed to minimize this loss. Adam adapts the learning rate for each parameter by combining the benefits of momentum and RMSProp, which leads to faster and more stable convergence when navigating the complex, non-convex optimization landscape inherent in medical image segmentation. The optimization process is initiated with a standard initial learning rate of 0.001, ensuring the model achieves robust performance on clinically relevant structures while improving overall pixel accuracy.

IV. Experiment

A. Datasets

Our experiment obtained three public datasets and one abdominal ultrasound dataset of our own. These datasets are:

- The Open Kidney Ultrasound Dataset. We took 120 of them as the training set, 20 as the validation set and 11 as the test set. https://github.com/rsingla92/kidneyUS/blob/main/README.md
- US simulation and segmentation Dataset .We took 287 of them as the training set, 14 as the validation set, and 44 as the test set. https://www.kaggle.com/datasets/ignaciorlando/ussimandsegm?select=abdominal_US
- Xu's Annotated Ultrasound Liver images Dataset. We took 487 of them as the training set, 68 as the validation set, and 190 as the test set. https://www.kaggle.com/datasets/orvile/annotated-ultrasound-liver-images-dataset
- The experimental team's own dataset. We took 833 images as the training set, 168 images as the validation set, and 64 images as the test set. The information of the datasets is shown in Table II.

TABLE II. THE INFORMATION OF THE DATASETS CONTAINED IN OUR EXPERIMENT

Dataset	Training images	Validation images	Test images	Resolution
Open Kidney	120	20	11	multiple
US simulation	287	14	44	640*480
Xu's	487	68	190	multiple
private data	833	168	64	multiple

B. Results

The following are the training and validation results of datasets on four different models. They are nnUNet, ResUNet, UNet, and our SPA-DCN-NET. The handheld probe (type: CPL) from Guangdong SonoStar in China was used to collect our private data, with a harmonic of 5.0Mhz and a focusing depth of 60mm. We labeled the liver regions of all datasets with LABELME 5.5 for 2-classification. A mixed dataset included the above three public datasets and our private dataset. There were a total of 1727 images as the training set. We use two schemes for training. First, we apply the dataset for each algorithm respectively (Table III). Second, we mix each training set, the compound weighted file is applied to each test set, and the calculated mean IoU in each dataset (see Table IV). The parameters of the training are as follows: batch_size = 4, num_epochs = 200, learning_rate = 0.001. The training loss curve is illustrated in Fig. 6. We use an INTEL G6405 with 32GB of RAM, Windows 10, and an NVIDIA M40 with 24GB RAM. The test results of the four models on the four datasets are as follows. In Scheme 1, independent training set, the SPA-DCN-NET model is stronger than all others (see Table III). In Scheme 2, integrated training set, SPA-DCN-NET is stronger in three tests (see Table IV). The results are shown in the tables.

From the data, we can see that our model has a good feature extraction ability for ultrasound images. Gated LCN makes the model more adaptable to ultrasound examination scenarios with unstable brightness. However, if the training data is too simple, such as the Open Kidney dataset, the model may only learn a fixed pattern, for example, always closing the LCN gate, thereby losing the ability to adapt dynamically. Therefore, to enable the gate control to learn when to switch, a certain degree of diversity of data is required. Fig. 7 provides a sample comparison of liver segmentation results in ultrasound between our SPA-DCN-NET and UNet/ResUNet/nnUNet in Scheme 2.

In the experiment, we evaluated the training data using several image brightness metrics, which are as follows: 1) Overall Uniformity Ratio U_1, 2) General Uniformity Ratio U_0, 3) Coefficient of Variation CV, 4) The average brightness/pixel value of the image E_{avg}. The IoU results of different segmentation algorithms and other metrics are presented in Table V. The samples of experimental results show that UNet/nnUNet/ResUNet is sensitive to noise and recognizes it as an organ boundary. When the E_{avg} is low (dark image), neurons cannot be activated, while our SPA-DCN-NET is still available. The effectiveness analysis results are shown in Table VI.

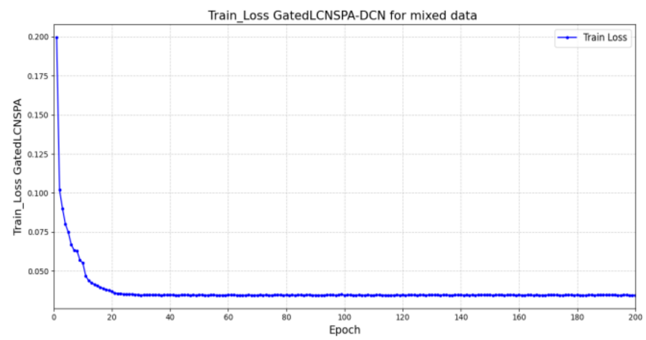

Fig. 6. Training loss of SPA-DCN-NET on mixed data.

TABLE III. SCHEME 1: THE MEAN IOU OF THE INDEPENDENT TRAINED ON EACH DATASET

	Open Kidney	US simulation	Xu's	Private data
UNet	0.8332	0.7402	0.7365	0.8555
nnUNet	0.8334	0.7414	0.7274	0.8560
ResUNet	0.8296	0.7339	0.7331	0.8558
SPA-DCN-NET	**0.8352**	**0.7557**	**0.7385**	**0.8594**

TABLE IV. SCHEME 2: THE MEAN IOU OF THE INTEGRATED TRAINING SET ON EACH DATASET

	Open Kidney	US simulation	Xu's	Private data
UNet	0.8092	0.7371	0.3038	0.8576
nnUNet	0.8243	0.7373	0.7361	0.8541
ResUNet	**0.8328**	0.7361	0.7253	0.8487
SPA-DCN-NET	0.8203	**0.7852**	**0.7710**	**0.8714**

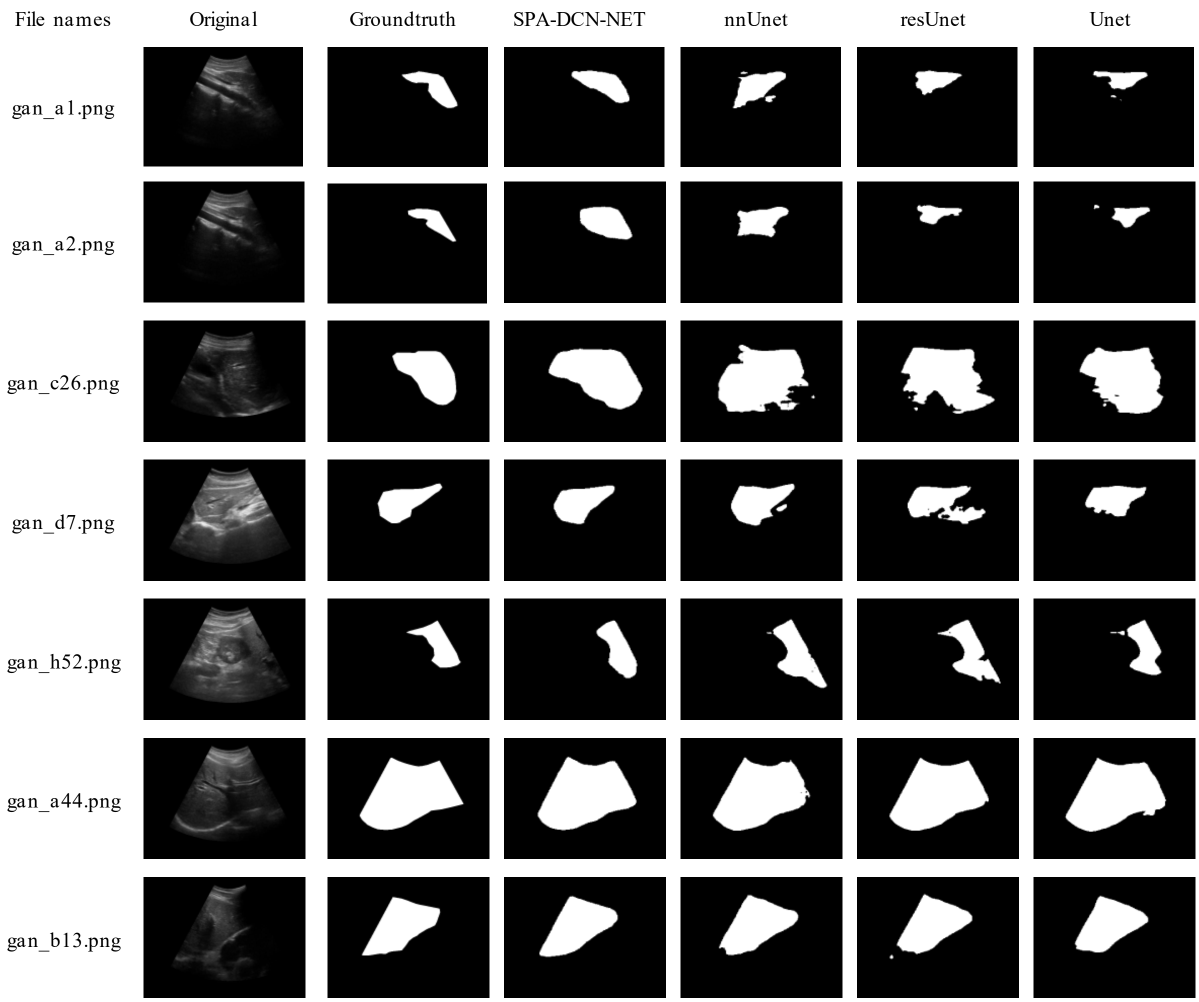

Fig. 7. A sample comparison of liver segmentation results between our SPA-DCN-NET and UNet/ResUNet/nnUNet.

TABLE V. COMPARISON OF SEGMENTATION IOU SCORES ACROSS DIFFERENT ALGORITHMS FOR SAMPLE IMAGES WITH BRIGHTNESS AND UNIFORMITY METRICS

File Name	U_1	U_0	E_{avg}	CV	nnUNet IoU	ResUNet IoU	UNet IoU	SPA-DCN-NET IoU
gan_a1.png	0.0047	0.035	28.58	1.1968	0.1444	0.1619	0.2610	0.6533
gan_a2.png	0.0049	0.0368	27.16	1.2741	0.1377	0.1516	0.2980	0.4935
gan_a44.png	0.0053	0.0245	40.79	0.7558	0.9460	0.9628	0.9437	0.9536
gan_b13.png	0.0055	0.0394	25.4	0.9994	0.8759	0.8815	0.8774	0.9287
gan_c26.png	0.0053	0.0312	32.09	0.7827	0.3863	0.4799	0.5345	0.6055
gan_d7.png	0.004	0.0134	74.58	0.6533	0.5449	0.7228	0.5003	0.7317
gan_h52.png	0.0045	0.0157	63.71	0.5736	0.5745	0.5967	0.6666	0.7441

U_0:General Uniformity Ratio; U_1:Overall Uniformity Ratio; E_{avg}: Average Brightness; CV: Coefficient of Variation.

TABLE VI. THE EFFECTIVENESS ANALYSIS OF SPA-DCN-NET

Characteristics	UNet / nnUNet / resUNet	SPA-DCN-NET	Comparison of results
Response to high CV (noise)	Sensitive. Misjudge the noise as the boundary.	Robust. Suppress the high-frequency noise area through the gating mechanism.	When gan_a1.png (CV=1.19) was present, the IoU gap was huge (0.14 vs 0.65).
Response to low Eavg (dark image)	Sensitive. The feature activation value is too low to activate neurons.	Robust. Local normalization focuses on relative contrast rather than absolute brightness.	SPA-DCN-NET can still maintain a relatively high segmentation accuracy in the dark image.

C. Ablation Experiment

We attempted four combinations: LCN, without LCN (SPA-DCN only), LCN before SPA followed by DCN input (LCNSPA-DCN), LCN after SPA followed by DCN input (SPA-LCNDCN), and gated LCN before SPA followed by DCN input (Gated LCNSPA-DCN). The gated LCNSPA-DCN performs best on the GAN dataset and our private dataset. The comparison result of IoU is shown in Table VII.

TABLE VII. ABLATION EXPERIMENT COMPARISON RESULT OF IOU ON SCHEME 2

	Open Kidney	US simulation	Xu's	Private data
SPA-DCN only	0.8282	0.7643	**0.7797**	0.8701
LCNSPA-DCN	**0.8338**	0.7646	0.7635	0.8517
SPA-LCNDCN	0.8316	0.7683	0.7613	0.8567
GATED LCNSPA-DCN	0.8203	**0.7852**	0.7710	**0.8714**

D. Analysis

We tested the performance of SPA-DCN-NET on four different sources of ultrasound liver datasets. SPA-DCN-NET demonstrated superior segmentation performance compared to well-known models (UNet, ResUNet, nnUNet) under two training strategies. Especially in the independent training Scheme (Scheme 1), the mean IoU of SPA-DCN-NET on all four datasets reached the highest value (83.52%, 75.57%, 73.85%, and 85.94% respectively), indicating that it has good generalization ability and adaptability to different data distributions. The introduction of the gated local contrast normalization (Gated LCN) mechanism significantly enhances the robustness of the model against interference factors such as unstable brightness, low contrast, and speckle noise in ultrasound images. Especially in Scheme 2, when the model was trained on mixed data and applied to each independent test set, SPA-DCN-NET significantly outperformed other models in mean IoU on the US simulation, Xu's and our private dataset. This indicates that the gating mechanism has successfully achieved a dynamic balance between LCN-enhanced robustness and detail retention of original features.

However, when the training data is less diverse and the number of images is small (such as the Open Kidney dataset), the gating mechanism may degenerate into a fixed pattern, for example, always closing the LCN branch, thereby losing its adaptive adjustment ability. This also explains why the performance of SPA-DCN-NET on Open Kidney (82.03%) is slightly lower than that of nnUNet (82.43%) in training Scheme 2. This indicates that the effectiveness of the gating mechanism depends on the diversity of the training data. Only on data with rich illumination, contrast, and deformation changes can the model truly learn when to enable LCN. Through four combinations of ablation experiments (no LCN, LCN pre-placement, LCN post-placement, Gated LCN pre-placement), the results showed that the Gated LCNSPA-DCN combination performed best on most datasets, especially on private datasets and US simulation datasets. This verifies the effectiveness of the design of introducing a Gated LCN in front of the SPA module. It not only retains the multi-scale attention's focusing ability on key areas but also enhances the stability of features in response to illumination changes through the LCN.

V. DISCUSSION

Although gated local contrast normalization has significant advantages in improving the model's robustness to the luminance instability and low contrast of ultrasound images, its introduction also brings several potential problems, mainly reflected in the following three aspects:

1) Dependency on training data diversity. The generalization ability and effectiveness of the gating mechanism are highly dependent on the training data's diversity in terms of illumination contrast and noise patterns. As shown in the experiments in the text, on the single Open Kidney dataset, the model tends to learn a fixed gating strategy, resulting in always closing the LCN branch, thereby losing its dynamic adaptability. This means that if the training set lacks a sufficiently diverse range of luminance variation samples, the gating mechanism may degenerate into a redundant module, or even be inferior to standard LCN or direct input of original features.

2) Increase the model complexity and training of instability. Gated LCN introduces additional learning parameters, such as two 1 x 1 convolution layer + sigmoid gating, and with the use of SPA, DCN cascade, a significant increase in the computational workload, especially in the superficial layer in MobileNetV3, deploying the module will lead to burst or vanishing gradients and overfitting. Furthermore, the training process exhibits oscillations or slow convergence.

3) Although the coupling mode of SPA and DCN enhances the perception ability of multi-scale context, its separation of guidance and execution design may lead to the misalignment of the focus area of attention with the actual sampling points, especially in high-noise areas, which are prone to misclassification. Secondly, although MobileNetV3 is adopted as the lightweight backbone, the computation of the SPA and DCN modules is relatively high. If deployed across all layers, it is prone to cause unstable training. Therefore, enabling this module only in deep, semantically rich regions reflects a fine trade-off between efficiency and accuracy. In the future, lightweight SPA structures can be explored, such as introducing depth-separable convolution, to achieve wider deployment. Finally, the currently adopted cross-entropy loss function has limitations in scenarios with severe category imbalance. In the early stage of the model, it is easily dominated by background pixels and converges slowly. Although SPA indirectly alleviates this problem, the introduction of Loss functions that are more suitable for medical segmentation, such as Generalized Dice Loss, may further improve the performance of small target segmentation.

VI. Conclusion

This study successfully addressed a critical challenge in medical image analysis by introducing SPA-DCN-NET, a novel framework designed to overcome the brightness instability and non-rigid geometric deformation in the segmentation of ultrasound images. Our main theoretical contribution is the proposed synergistic adaptation mechanism. We combine gated local contrast normalization, spatial pyramid attention, and a deformable convolutional network together, establishing an innovative, end-to-end learnable loop wherein illumination and geometry mutually inform one another. This enhances the robustness of traditional segmentation algorithms against brightness variations and geometric deformation. In the experimental results, our SPA-DCN-NET performance surpassed benchmark models such as UNet, nnUNet and ResUNet on four different datasets, especially on the private datasets. It achieved an average IoU as high as 85.94%, demonstrating its potential in real-world applications. However, this study mainly focuses on binary segmentation of the liver; the multi-category recognition of blood vessels and lesions still needs to be explored. Future research will aim to adapt the architecture for multi-organ and multi-lesion segmentation tasks. Further optimizing its computational efficiency to adapt to real-time, portable ultrasound diagnosis, thereby enhancing its clinical practicality.

Acknowledgment

We gratefully acknowledge Sonostar Ltd., ShenZhen, China. Without their equipment and technical support, the experiment cannot carry on. Thanks to Prof. Afizan, who offered valuable suggestions for the study. And all my families, they supported me behind my back.

References

[1] Younossi, Z. M., Wong, G., Anstee, Q. M., & Henry, L. (2023). The global burden of liver disease. Clinical Gastroenterology and Hepatology, 21(8), 1978-1991.

[2] Jiang, J., Jiang, X., Xu, L., Zhang, Y., Zheng, Y., & Kong, D. (2023). Noise-robustness test for ultrasound breast nodule neural network models as medical devices. Frontiers in Oncology, 13, 1177225.

[3] Ronneberger, O. , Fischer, P., & Brox, T. . (2015). U-net: convolutional networks for biomedical image segmentation. Springer, Cham.

[4] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: Self-adapting framework for u-net-based medical image segmentation. Methods , 194, 40-48.

[5] Sivaraman, K., & Murthy, A. (2018). Object Recognition under Lighting Variations using Pre-Trained Networks. 2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1-7. https://doi.org/10.1109/aipr.2018.8707399.

[6] Sietzen, S., Lechner, M., Borowski, J., Hasani, R., & Waldner, M. (2021). Interactive Analysis of CNN Robustness. Computer Graphics Forum, 40. https://doi.org/10.1111/cgf.14418.

[7] Hu, C., Sapkota, B., Thomasson, J., Bagavathiannan, M., Nez-Guanter, J., & Jar, D. (2021). Influence of Image Quality and Light Consistency on the Performance of Convolutional Neural Networks for Weed Mapping. Remote. Sens., 13, 2140. https://doi.org/10.3390/rs13112140.

[8] Petkovic, J., & Fioresi, R. (2024). Spontaneous Emergence of Robustness to Light Variation in CNNs With a Precortically Inspired Module. Neural Computation, 36, 1832-1853. https://doi.org/10.1162/neco_a_01691.

[9] Jain, A., Dubey, S., Singh, S., Santosh, K., & Chaudhuri, B. (2024). Non-Uniform Illumination Attack for Fooling Convolutional Neural Networks. ArXiv, abs/2409.03458. https://doi.org/10.48550/arxiv.2409.03458.

[10] Jian, M., Qi, Q., Dong, J., Yin, Y., & Lam, K. M. (2018). Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. Journal of visual communication and image representation, 53, 31-41.

[11] Al Qurri, A., & Almekkawy, M. (2023). Improved UNet with attention for medical image segmentation. Sensors, 23(20), 8589.

[12] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 764-773).

[13] Wang, D., & Lan, J. (2021). A Deformable Convolutional Neural Network with Spatial-Channel Attention for Remote Sensing Scene Classification. Remote Sensing, 13(24), 5076. https://doi.org/10.3390/rs13245076

[14] Guo, J., Ma, X., Sansom, A., McGuire, M., Kalaani, A., Chen, Q., ... & Fu, S. (2020, July). Spanet: Spatial pyramid attention network for enhanced image recognition. In 2020 IEEE international conference on multimedia and expo (ICME).

[15] Howard, A. , Sandler, M. , Chen, B. , Wang, W. , Chen, L. C. , & Tan, M. , et al. (2020). Searching for mobilenetv3. IEEE

[16] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[17] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Springer, Cham.

[18] Oktay, O., Schlemper, J., Le Folgoc, L., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

[19] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[20] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In International conference on machine learning (pp. 1243-1252). PMLR.

[21] Rahman, H., Bukht, T. F. N., Imran, A., Tariq, J., Tu, S., & Alzahrani, A. (2022). A Deep Learning Approach for Liver and Tumor Segmentation in CT Images Using ResUNet. Bioengineering, 9(8), 368. https://doi.org/10.3390/bioengineering9080368

[22] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[23] Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., & Nevatia, R. (2020, August). SPAN: Spatial pyramid attention network for image manipulation localization. In European conference on computer vision (pp. 312-328). Cham: Springer International Publishing.

[24] Li, C., Du, D., Zhang, L., Wen, L., Luo, T., Wu, Y., & Zhu, P. (2020, August). Spatial attention pyramid network for unsupervised domain adaptation. In European Conference on Computer Vision (pp. 481-497). Cham: Springer International Publishing.

[25] Yu, Y., Zhang, Y., Cheng, Z., Song, Z., & Tang, C. (2024). Multi-scale spatial pyramid attention mechanism for image recognition: An effective approach. Engineering Applications of Artificial Intelligence, 133, 108261.

[26] Rad, M., Roth, P. M., & Lepetit, V. (2020). ALCN: Adaptive local contrast normalization. Computer Vision and Image Understanding, 194, 102947.

[27] Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. Visual neuroscience, 9(2), 181-197.

[28] Ortiz, A., Robinson, C., Morris, D., Fuentes, O., Kiekintveld, C., Hassan, M. M., & Jojic, N. (2020). Local context normalization: Revisiting local normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11276-11285).

[29] Chen, J., Shen, X., Zhao, Y., Qian, W., Ma, H., & Sang, L. (2024). Attention gate and dilation U-shaped network (GDUNet): an efficient breast ultrasound image segmentation network with multiscale information extraction. Quantitative Imaging in Medicine and Surgery, 14(2), 2034.

[30] Zeng, Y., Tsui, P. H., Wu, W., Zhou, Z., & Wu, S. (2021). Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net. Journal of Digital Imaging, 34(1), 134-148.

[31] Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., ... & Rueckert, D. (2018). Attention-gated networks for improving ultrasound scan plane detection. arXiv preprint arXiv:1804.05338.

[32] Zhu, X., Cheng, H., Zhang, Q., Tu, S., & Xiang, T. (2020). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10339-10348).

[33] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.