

---

# A SEMANTIC SEGMENTATION ALGORITHM FOR PLEURAL EFFUSION BASED ON DBIF-AUNET \*

---

Ruixiang Tang<sup>1</sup>, Mingda Zhang<sup>1</sup>, Jianglong Qin<sup>1,2\*</sup>, Yan Song<sup>3,4</sup>, Yi Wu<sup>4</sup>, Wei Wu<sup>5</sup>

<sup>1</sup>School of Software, Yunnan University, Kunming 650500

<sup>2</sup>Yunnan Provincial Key Laboratory of Software Engineering,  
Yunnan University, Kunming 650500

<sup>3</sup>Department of Information Technology, Army Medical University, Chongqing 400038

<sup>4</sup>Division of Digital Medical Education & Research, Department of Biomedical  
Engineering and Imaging Medicine, Army Medical University, Chongqing 400038

<sup>5</sup>Department of Thoracic Surgery, Southwest Hospital, Chongqing 400038

\*Corresponding author: qinjianglong@ynu.edu.cn

## ABSTRACT

Pleural effusion semantic segmentation can significantly enhance the accuracy and timeliness of clinical diagnosis and treatment by precisely identifying disease severity and lesion areas. Currently, semantic segmentation of pleural effusion CT images faces multiple challenges. These include similar gray levels between effusion and surrounding tissues, blurred edges, and variable morphology. To address these challenges, we propose the Dual-Branch Interactive Fusion Attention model (DBIF-AUNet). This model constructs a densely nested skip-connection network and innovatively refines the Dual-Domain Feature Disentanglement module (DDFD). The DDFD module orthogonally decouples the functions of dual-domain modules to achieve multi-scale feature complementarity and enhance characteristics at different levels. Concurrently, we design a Branch Interaction Attention Fusion module (BIAF) that works synergistically with the DDFD. This module dynamically weights and fuses global, local, and frequency band features, thereby improving segmentation robustness. Furthermore, we implement a nested deep supervision mechanism with hierarchical adaptive hybrid loss to effectively address class imbalance. Through validation on 1,622 pleural effusion CT images from Southwest Hospital, DBIF-AUNet achieved IoU and Dice scores of 80.1% and 89.0% respectively. These results outperform state-of-the-art medical image segmentation models U-Net++ and Swin-UNet by 5.7%/2.7% and 2.2%/1.5% respectively, demonstrating significant optimization in segmentation accuracy for complex pleural effusion CT images.

**Keywords** Pleural effusion · Densely nested skip connections network · Dual-Domain Feature Disentanglement module · Branched-Interactive Attention Fusion module · Nested deep supervision

## 1 Introduction

Pleural effusion is a common clinical condition resulting from various etiologies, including heart failure, infection, and malignant tumors, which lead to fluid accumulation in the pleural cavity between the parietal and visceral pleura [1]. Accurate identification and segmentation of pleural effusion are crucial for assessing its severity and guiding therapeutic interventions. Currently, traditional pleural effusion identification relies primarily on manual interpretation, which suffers from low efficiency, strong subjectivity, and limited sensitivity in detecting small amounts of effusion [2]. These limitations are particularly evident in mesothelioma diagnosis, where diagnostic accuracy and sensitivity remain suboptimal [3].

In recent years, deep learning technology has achieved significant progress in medical image semantic segmentation [4]. Methods based on convolutional neural networks (CNNs) have demonstrated strong performance across various

---

\**Citation:* Tang et al. A Semantic Segmentation Algorithm for Pleural Effusion Based on DBIF-AUNet. 2025.

medical image segmentation tasks [5]. However, pleural effusion presents unique challenges in CT images, as it often displays similar or gradually transitioning gray values with lung parenchyma, diaphragm, chest wall, and other structures. This similarity causes traditional segmentation methods to fail, especially when effusion coexists with atelectasis and pleural thickening, which further reduces edge discrimination. Moreover, small amounts of effusion (<300ml) may only appear as blurred posterior costophrenic angles in supine CT, while large amounts can occupy the entire chest cavity and change from crescent to circular shapes. This morphological variability significantly increases the generalization difficulty for segmentation models. Therefore, precise segmentation remains a considerable challenge. To address these limitations, this paper proposes a network based on dual-domain module decoupling and branch interactive attention fusion (Dual-Branch Interactive Fusion Attention U-Net, DBIF-AUNet) for semantic segmentation of pleural effusion. This model leverages the synergistic advantages of dual-domain feature decoupling and multi-branch interactive attention mechanisms to better handle complex pleural effusion CT images.

## 2 Related Work

### 2.1 U-shaped Deep Learning Networks

U-shaped deep learning networks, particularly U-Net and its variants, have been extensively studied and successfully applied in biomedical image segmentation. U-Net was first proposed by Olaf Ronneberger et al. in 2015 [6], featuring a contracting path (encoder) and a symmetric expanding path (decoder) that form a distinctive U-shaped structure. The contracting path captures contextual information, while the expanding path enables precise localization. Skip connections allow the network to transfer feature maps from the contracting path to the symmetric expanding path, thereby better preserving image details. Building upon U-Net, Attention U-Net introduces attention mechanisms that help the network focus more effectively on important regions in the image, thus improving segmentation results [7].

U-Net++ further enhances feature reuse and fusion by introducing nested and dense skip connections [8]. To overcome the limitations of convolutional neural networks in learning global and long-range semantic information interactions, Hu Cao et al. proposed Swin-UNet, which integrates a Transformer-based architecture into the U-Net structure [9]. However, skip connections in traditional U-shaped deep learning networks directly concatenate shallow and deep features, inevitably leading to semantic gaps and detail loss. To address this issue, we propose a cross-level feature aggregation network with densely nested skip connections, which significantly improves multi-scale feature fusion efficiency through cross-level feature transfer via skip connections.

### 2.2 Dual-Domain Feature Fusion

Traditional semantic segmentation methods primarily focus on spatial domain features of images, namely spatial relationships between pixels and local texture information [10]. To address the limitations of single-domain approaches, researchers have begun exploring the use of features from other domains, particularly the frequency domain, to assist semantic segmentation. Frequency domain features in dual-domain modules can provide global structure and texture information of images, effectively complementing spatial domain features and thus improving segmentation robustness [11]. While spatial domain features describe local texture and spatial relationships of images, frequency domain features capture global structure and frequency information [11, 12]. By fusing spatial and frequency domain features, models can simultaneously focus on both local details and global structure of images, thereby improving segmentation accuracy. Dual-domain feature fusion has achieved significant results in multiple semantic segmentation tasks, including remote sensing image segmentation, medical image segmentation, and autonomous driving scene segmentation [10, 11]. However, current traditional dual-domain methods only enhance global features through frequency domain channel attention without explicitly decoupling structure and texture information, resulting in feature confusion problems. We orthogonally decouple traditional dual-domain feature fusion modules and design a dual-domain feature disentanglement module to effectively handle complexity and edge blurriness in images while achieving multi-scale feature complementarity.

### 2.3 Attention Mechanisms in Medical Image Segmentation

In deep learning, attention mechanisms enable models to focus on the most important features in input data by learning weight distributions, thereby improving model performance and efficiency [13]. These mechanisms have been widely applied in medical image analysis with remarkable success. In lung CT image semantic segmentation, Yang et al. designed AttGGO-Net, which utilizes self-attention and cross-attention mechanisms to improve segmentation accuracy [14]. Chen et al. proposed a robust Attention-Guided and Noise-Resistant framework (AGNR), which integrates attention-guided mechanisms and noise resistance learning to enhance segmentation accuracy [15].

For multi-scale attention mechanism fusion, SMANet employs superpixel-guided multi-scale attention, dividing medical images into different regions through superpixel segmentation and using multi-scale attention mechanisms to fuse features from these regions, achieving more precise segmentation [16]. To address the problem that traditional attention mechanisms tend to over-focus on globally most salient effective features while suppressing secondary salient features, Zhan et al. proposed FSA-Net, which improves medical image segmentation by releasing global suppression information [17]. However, current attention mechanisms struggle to balance global structure, local details, and frequency domain features when dealing with complex edges, and static weight allocation lacks dynamic adaptability. We propose a multi-branch interactive attention fusion module that employs three-level feature information processing and gated interactive selection mechanisms, significantly enhancing the ability to delineate complex edges and overcoming the locality limitations of traditional single attention.

## 2.4 Deep Supervision

Deep supervision technology has been widely applied in semantic segmentation by introducing intermediate layer supervision in neural network training to strengthen learning signals and overcome gradient vanishing problems in deep network training [18, 19]. Due to the complexity of deep networks and their dependence on large amounts of annotated data, the training process often faces problems such as gradient vanishing or convergence difficulties. Deep supervision technology improves model training efficiency and performance by optimizing intermediate layer feature representations and transmitting supervision signals more directly to intermediate layers of the network. In medical image segmentation, deep supervision has significantly improved segmentation accuracy for various pathologies [20, 21, 22], while simultaneously accelerating network convergence speed and improving final model performance by providing additional learning signals [23]. However, current deep supervision approaches (such as UNet++) [8] use identical loss functions at different levels, ignoring semantic differences between levels and struggling with class imbalance problems. We design a nested deep supervision mechanism to address these issues, effectively improving segmentation accuracy for small targets and alleviating gradient vanishing problems.

## 3 Method

### 3.1 Nested Skip Connection Structure of DBIF-AUNet

We propose DBIF-AUNet (Dual-Branch Interactive Fusion Attention U-Net) for handling images with complex and diverse edges. DBIF-AUNet adopts a nested skip connection U-shaped network architecture similar to UNet++ [8] and ANUNet [24]. Figure 1 illustrates the densely nested skip connection structure of DBIF-AUNet.

The improved densely nested skip connection structure consists of five essential parts: the DBIF-AUNet encoder, decoder, DDFD dual-domain feature fusion module, BIAF branch interactive attention fusion module, and nested deep supervision module.

As shown in Figure 1, cross-level skips first aggregate deep features and shallow features with current layer features. Note that this figure only displays the outermost skip connection structure, while the internal skip connection method follows the same pattern as the outermost. Subsequently, the input multi-level features are processed through the DDFD module with functionally decoupled three independent sub-modules to obtain deep semantics, current layer refined features, and shallow details. To avoid information loss caused by feature transfer in the decoder in traditional connection methods, the DDFD module directly transfers nested features to the three input branches of the BIAF module, which then processes feature information from different layers for interactive dynamic fusion.

### 3.2 Dual-Domain Feature Disentanglement Module

To address the structural complexity of pleural effusion images, we select dual-domain modules that can combine spatial and frequency domains to enhance feature information. To solve the confusion limitations of traditional dual-domain fusion, we design a module decoupling enhancement method for multi-level features, performing targeted processing on features at different levels. The dual-domain module is functionally orthogonally decomposed into a dual-domain decoupling module, which strengthens boundary perception through spatial strip attention and combines with frequency domain band attention decoupling. Subsequently, it achieves multi-scale feature complementarity through bidirectional frequency band modulation and spatial-frequency domain alignment.

As shown in Figure 2, the deep, shallow, and current layer features input from skip connections are processed through the DDFD (Dual-Domain Feature Disentanglement) module after feature aggregation and normalization. For the shallow features of layer l0, these consist of shallow convolution structures with 3×3 convolution kernel plus ReLU activation. The three dual-domain module decoupling functional modules in the DDFD module explicitly separate

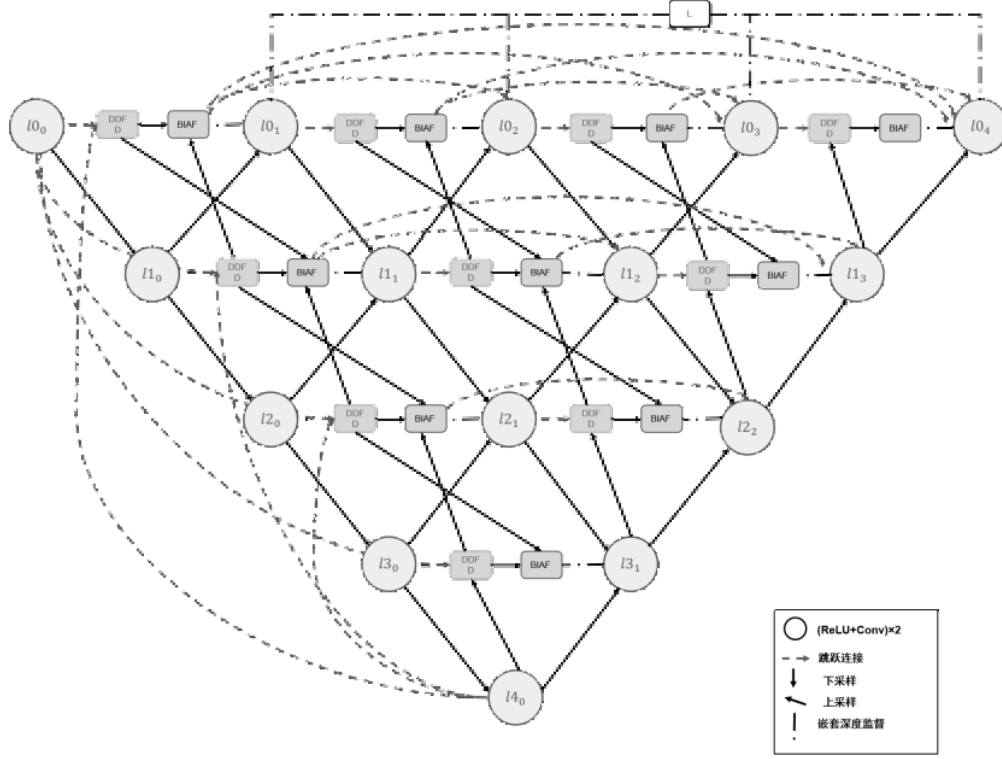


Figure 1: Nested skip connection structure of DBIF-AUNet

semantic information at different levels. Specifically, the global branch captures the overall morphology of the image, the local branch strengthens edge gradient changes, and the channel branch filters noise and enhances cross-channel texture expression. The processing methods for each branch are as follows:

1. Global average pooling (GAP) is used to extract low-frequency global structure of deep features, and after enhancing cross-channel interaction through frequency domain channel attention, global context features with frequency bands are obtained and input to the global branch through bilinear interpolation.

$$\mathbf{c} = \sigma \left( \mathbf{W}_2 \cdot \text{ReLU} \left( \mathbf{W}_1 \cdot \left| \mathcal{DCT} \left( \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_{i,j} \right) \right| + \mathbf{b}_1 \right) + \mathbf{b}_2 \right) \odot \mathcal{DCT} \left( \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_{i,j} \right) \quad (1)$$

where  $\sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_{i,j}$  represents calculating spatial sum and output vector independently for each channel,  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are learnable parameters,  $\mathcal{DCT}$  is 2D DCT transform,  $\text{ReLU}$  and  $\sigma$  are activation functions, and  $\odot$  denotes element-wise multiplication.

2. Boundary perception is strengthened through Strip Pooling, which directly associates shallow detail features with deep semantics and combines spatial position weights with frequency domain energy distribution. After performing 2D discrete wavelet transform (DWT) on input feature  $\mathbf{X}$  through wavelets, we separate low-frequency approximation component  $\mathbf{X}_{LL}$  and high-frequency detail component  $\mathbf{X}_{HH}$ . Meanwhile, Gabor filters balance spatial resolution and frequency band resolution, achieving spatial skip attention and spatial-frequency domain balance mechanism to obtain high-frequency edge features. The fusion features are shown in Equation 5, and the results are finally input to the local branch after spatial-frequency domain alignment.

$$\mathbf{y}_{c,i}^h = \frac{1}{W} \sum_s^W \mathbf{X}_{c,i,j} \in \mathbb{R}^{C \times H \times 1} \quad (2)$$

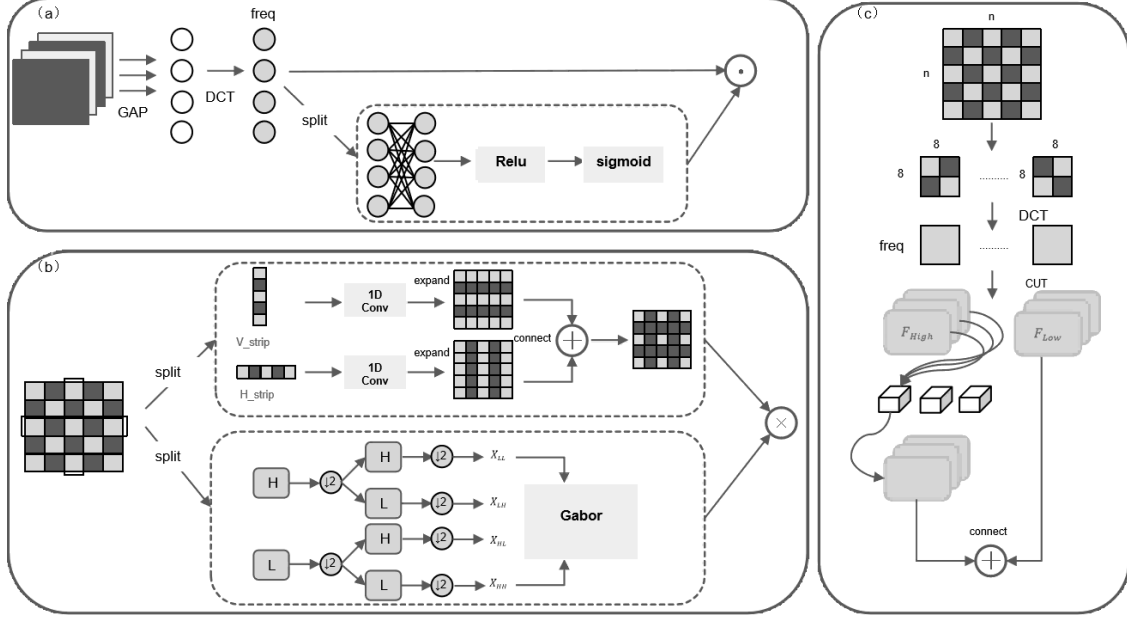


Figure 2: Decoupling functional structure of the dual-domain feature disentangling module

$$\mathbf{y}_{\mathbf{c}\mathbf{j}}^{\mathbf{v}} = \frac{1}{\mathbf{H}} \sum_{\mathbf{i}=1}^{\mathbf{H}} \mathbf{X}_{\mathbf{c},\mathbf{i},\mathbf{j}} \in \mathbb{R}^{\mathbf{C} \times \mathbf{1} \times \mathbf{W}} \quad (3)$$

$$\mathbf{G}_{\theta,\mathbf{f}}(\mathbf{i},\mathbf{j}) = \exp\left(-\frac{\mathbf{i}'^2 + \gamma^2 \mathbf{j}'^2}{2\sigma^2}\right) \cos(2\pi \mathbf{f} \mathbf{i}' + \phi) \quad (4)$$

where  $\mathbf{i}' = \mathbf{i} \cos \theta + \mathbf{j} \sin \theta$  and  $\mathbf{j}' = -\mathbf{i} \sin \theta + \mathbf{j} \cos \theta$ . Here, orientation parameter  $\theta$  and frequency  $\mathbf{f}$  adaptively adjust spatial and frequency band resolution weights.

$$\hat{\mathbf{X}}_{\text{local}} = \text{Concat}(\mathbf{y}^{\mathbf{h}}, \mathbf{y}^{\mathbf{v}}) \otimes \mathbf{G}_{\theta,\mathbf{f}}(\mathbf{X}_{\text{LL}}, \mathbf{X}_{\text{HH}}) \quad (5)$$

where  $\otimes$  represents spatial weighting operation, enhancing collaborative expression of edges and textures.

3. Through 2D DCT transform as shown in Equation 6, we separate low-frequency (structural information) and high-frequency (texture details) of current layer features, along with cross-band texture obtained from frequency band separation. These are then reconstructed and input to the channel branch through DCT-IDCT.

$$\mathbf{F}(\mathbf{u}, \mathbf{v}) = \alpha(\mathbf{u})\alpha(\mathbf{v}) \sum_{\mathbf{i}=0}^{\mathbf{H}-1} \sum_{\mathbf{j}=0}^{\mathbf{W}-1} \mathbf{X}_{\mathbf{c},\mathbf{i},\mathbf{j}} \cos\left[\frac{\pi(2\mathbf{i}+1)\mathbf{u}}{2\mathbf{H}}\right] \cos\left[\frac{\pi(2\mathbf{j}+1)\mathbf{v}}{2\mathbf{W}}\right] \quad (6)$$

where  $\alpha(u) = \sqrt{\frac{1}{H}}$  when  $u=0$ , or  $\sqrt{\frac{2}{H}}$  when  $u>0$ .

### 3.3 Branch Interactive Attention Fusion Module

Due to the variable morphology of pleural effusion in CT images, such as diffuse small effusion and encapsulated complex effusion, along with blurred boundaries with surrounding tissues including lung parenchyma and pleura, we design a module capable of multi-level feature fusion and interactive weight adjustment. This module builds upon traditional SE attention modules and combines with feature information processed by the dual-domain decoupling module. The module consists of three components: branch parallel enhancement module, branch attention fusion module, and interactive attention fusion module for weighted feature fusion, which dynamically adjusts weights

interactively for different morphologies of complex images. After aligning the feature channels from Section 2.2 and inputting them into the three input branches of this module, the module adopts a three-level architecture of parallel processing, followed by feature fusion, and finally dynamic weighting. By processing features from three scales - global branch for low-frequency structure, local branch for high-frequency edges, and channel branch for frequency band texture - and then performing dynamic interactive fusion, it achieves balanced expression of structure and details in complex pleural effusion CT images, thus enhancing robustness in segmentation environments with variable morphology and blurred boundaries. Figure 3 shows the structure of the BIAF (Branched-Interactive Attention Fusion Module).

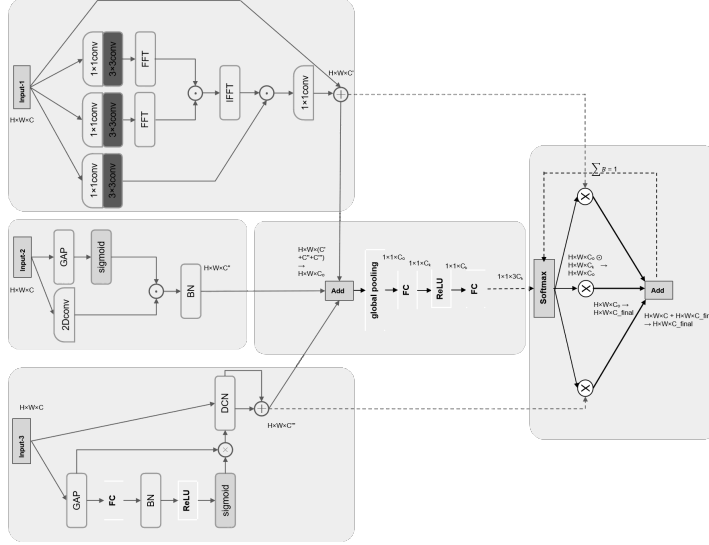


Figure 3: Branch Interactive Attention Fusion Module

The branch fusion module further processes the input feature information before fusion by adopting parallel enhancement of frequency domain-guided global modeling strengthening, local deformation adaptive enhancement, and channel-aware noise filtering. The processing procedure is shown in the following formulas:

The global branch uses frequency domain self-attention to transform features to frequency domain through FFT, establishing long-range dependencies and solving the locality limitations of spatial convolution. Meanwhile, cross-band complementarity strengthens the interaction from low-frequency contour information to mid-frequency texture information to high-frequency edge details, avoiding loss of key information:

$$\mathbf{A}_g = \mathcal{CFC}(\text{Softmax}(\frac{\mathcal{F}(\mathbf{Q}) \cdot \mathcal{F}(\mathbf{K})^T}{\sqrt{\mathbf{d}_k}})) \quad (7)$$

where  $\mathcal{CFC}()$  denotes the cross-band correlation function.

The local branch adopts local feature adaptive dynamic convolution groups that adaptively adjust convolution kernel parameters according to input content, enhancing robustness to deformation occlusion. Then it uses frequency band normalization to group and standardize features at different frequencies, solving training instability caused by frequency domain distribution differences:

$$\mathbf{A}_l = \frac{1}{K} \sum_{k=1}^K \text{Sigmoid}(\mathbf{U}_k \cdot \text{GAP}(X))(\mathbf{X} * \mathbf{W}_k) \oslash \sqrt{\sigma_k^2 + \epsilon} \quad (8)$$

where  $\sigma_k^2$  is the variance of the  $k$ -th frequency band,  $\oslash$  denotes element-wise division, and  $\mathbf{U}_k$ ,  $\mathbf{W}_k$  are convolution kernel parameters.

The channel branch uses deformable convolution to learn spatial offsets, adapting to non-rigid deformations, while adjusting channel soft thresholds through learnable thresholds to filter irrelevant frequency band noise:

$$\mathbf{A}_c = \eta \left( \sum_{\mathbf{n} \in \mathbb{N}} \mathbf{w}_n \cdot X(\mathbf{p}_0 + \mathbf{p}_n + \Delta \mathbf{p}_n) \right) \circ \eta \left( \sum_{\mathbf{m} \in \mathcal{M}} \mathbf{v}_m \cdot \mathbf{X}(\mathbf{q}_0 + \mathbf{q}_m + \Delta \mathbf{q}_m) \right) \quad (9)$$

where  $\eta(z) = \text{sign}(z) \cdot \max(0, |z| - \tau)$ ,  $p_n$  and  $q_m$  represent offsettable spatial variables, and  $\circ$  is the composite function.

Subsequently, dynamic fusion is performed through the branch attention fusion module:

$$\mathbf{Y} = \sigma(\mathbf{W}_{f_2}^\top \cdot \text{ReLU}(\mathbf{W}_{f_1}^\top \cdot \text{GAP}(Z))) + \text{Concat}(\mathbf{A}_g * \mathbf{X}_1, \mathbf{A}_1 * \mathbf{X}_2, \mathbf{A}_c * \mathbf{X}_3) \quad (10)$$

where  $\mathbf{W}_{f_1}^\top$  and  $\mathbf{W}_{f_2}^\top$  are fully connected weights.

Finally, three-branch decoupling is performed in the interactive attention fusion module, where weights are normalized through the Softmax function and interactive dynamic weighted output is performed as shown in the following formula:

$$\mathbf{B}_i = \frac{\exp(\mathbf{W}_{b_i}^\top \cdot Y)}{\sum_{j=1}^3 \exp(\mathbf{W}_{b_j}^\top \cdot Y)} \quad \text{for } i = 1, 2, 3 \quad (11)$$

where  $\mathbf{W}_{b_i}^\top$  represents channel attention weight.

$$\mathbf{O} = \sum_{i=1}^3 \mathbf{B}_i \odot \mathbf{Y}_i \quad (12)$$

Finally, interactive gated selection is performed for three-way information, where  $B_i$  denotes the weight vector.

### 3.4 Nested Deep Supervision

To implement the nested deep supervision structure designed in this paper, we develop a multi-level nested deep supervision hybrid loss. In addition to adding supervision nodes at the end of each upsampling path in the original UNet++ deep supervision mechanism, we also insert deep supervision points in each BIAF module output path, solving the multi-level supervision balance problem through adaptive weights. The calculation process of the loss function for this method proceeds as follows:

First, we calculate a hybrid loss at each supervision point using Dice Loss, Focal Loss, and Binary Cross-Entropy Loss respectively. These losses handle class imbalance and optimize segmentation boundaries, focus on hard samples, and provide stable gradients. The weighted hybrid loss function is shown in Equation 16.

$$\mathbf{L}_{\text{Dice}}^{(k)} = 1 - \frac{2 \sum (\mathbf{Y}_{\text{true}} \odot \mathbf{Y}_{\text{pred}}^{(k)}) + \epsilon}{\sum \mathbf{Y}_{\text{true}} + \sum \mathbf{Y}_{\text{pred}}^{(k)} + \epsilon} \quad (13)$$

where  $Y_{\text{true}}$  represents the ground truth label and  $Y_{\text{pred}}^{(k)}$  represents the prediction.

$$\mathbf{L}_{\text{Focal}}^{(k)} = -\frac{1}{N} \sum_{i=1}^N [\alpha(1 - \mathbf{Y}_{\text{pred},i}^{(k)})^\gamma \mathbf{Y}_{\text{true},i} \log(\mathbf{Y}_{\text{pred},i}^{(k)}) + (1 - \alpha)(\mathbf{Y}_{\text{pred},i}^{(k)})^\gamma (1 - \mathbf{Y}_{\text{true},i}) \log(1 - \mathbf{Y}_{\text{pred},i}^{(k)})] \quad (14)$$

$$\mathcal{L}_{\text{BCE}}^{(k)} = -\frac{1}{N} \sum_{i=1}^N [\mathbf{Y}_{\text{true},i}^{(k)} \log(\mathbf{Y}_{\text{pred},i}^{(k)}) + (1 - \mathbf{Y}_{\text{true},i}^{(k)}) \log(1 - \mathbf{Y}_{\text{pred},i}^{(k)})] \quad (15)$$

$$\mathcal{L}_{\text{hybrid}}^{(k)} = \lambda_{\text{Dice}} \mathcal{L}_{\text{Dice}}^{(k)} + \lambda_{\text{Focal}} \mathcal{L}_{\text{Focal}}^{(k)} + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}^{(k)} \quad (16)$$

where hyperparameter weights  $\lambda_{\text{Dice}}$ ,  $\lambda_{\text{Focal}}$ , and  $\lambda_{\text{BCE}}$  are set to 0.4, 0.3, and 0.3 respectively.

The supervision point function is defined as:

$$\mathbf{H}(\mathbf{X}) = \mathbf{Conv}_{1 \times 1} \circ \text{ReLU} \circ \mathbf{Conv}_{3 \times 3} \circ \text{Sigmoid}(X) \quad (17)$$

Upsampling endpoint supervision is calculated as:

$$\mathcal{L}_{\mathbf{U}_j} = \mathcal{L}_{\text{hybrid}}(\mathcal{H}(\mathbf{I}_{0j}), \mathbf{Y}_{\text{true}}) \quad (18)$$

BIAF output path loss is computed as:

$$\mathcal{L}_{\mathbf{B}_j} = \mathcal{L}_{\text{hybrid}}(\text{Deconv}_{2j}(\mathcal{H}(\mathbf{S}_j)), \mathbf{Y}_{\text{true}}) \quad (19)$$

Since different supervision points have varying importance, we assign different weights according to their layer depth, with weight factors as shown in Equation 21:

$$\omega_r^{(j)} = \frac{\text{Area}(\mathbf{X}_{0,j})}{\sum_{m=1}^4 \text{Area}(\mathbf{X}_{0,m})} = \frac{4^{j-1}}{\sum_{m=1}^4 4^{m-1}} \quad (20)$$

$$\mathbf{w}_{\mathbf{S}_j} = \text{Softmax}(1.5 \cdot \omega_r^{(j)} \cdot 0.7) \quad (21)$$

$$\mathbf{w}_{\mathbf{U}_j} = \text{Softmax}(1.5 \cdot \omega_r^{(j)} \cdot 0.9) \quad (22)$$

The final weighted multi-level supervision hybrid loss function is shown in Equation 24:

$$\mathcal{L}_{\text{Total}} = \sum_{j=1}^4 \mathbf{w}_{\mathbf{U}_j} \mathcal{L}_{\mathbf{U}_j} + \sum_{j=1}^4 \mathbf{w}_{\mathbf{S}_j} \mathcal{L}_{\mathbf{B}_j} \quad (23)$$

## 4 Experiments

### 4.1 Dataset

Due to the lack of high-quality public 2D pleural effusion CT image datasets, the dataset used in this paper comes from 9 expert-annotated patient imaging data containing various thoracic diseases authorized by Southwest Hospital. The dataset includes 1,622 DICOM format CT files and supporting Amira annotation files, with ethics approval number (B)KY2023062. All annotations were completed by thoracic surgeons and verified for consistency with a kappa value of 0.87, and sensitive privacy information was removed from the data. The dataset processing workflow is as follows:

**Standardization and Conversion:** We unified DICOM files with window width and level settings using lung window parameters of 1500/-600 HU, and standardized voxel spacing to 1.0×1.0×1.0 mm<sup>3</sup> through bilinear interpolation. All files were converted to 512×512 pixel 24-bit PNG format CT images, totaling 1,622 images. Simultaneously, we unified Amira annotation file effusion color to RGB values of 255,0,0, then normalized and converted them to 2D PNG format multi-class masks.

**Binary Extraction:** Based on HSV threshold settings with  $\text{H} \in [0, 10]$ ,  $\text{S} \in [200, 255]$ , and  $\text{V} \in [200, 255]$ , we extracted pleural effusion regions from images originally containing various thoracic diseases, generating single-channel binary masks with 0 for background and 255 for foreground.

**Data Pairing and Normalization:** We strictly aligned CT and mask filenames, linearly normalized CT gray values to the range [0,1], and binarized masks to values of 0 and 1.

**Stratified Dataset Division:** Using stratified sampling by patient ID to avoid cross-patient data leakage, we divided the dataset into a training set of 1,298 images (80

**Quality Control:** We randomly sampled 5

The processed dataset images are shown in Figure 4:



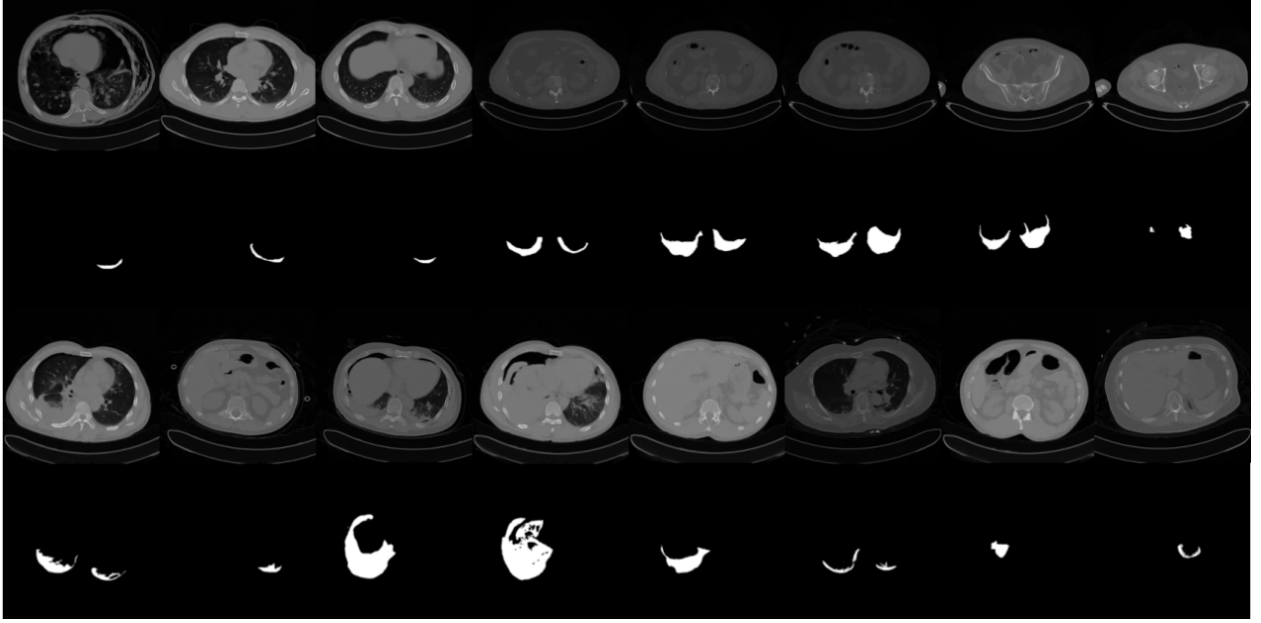


Figure 4: The constructed pleural effusion CT image dataset

#### 4.2 Experimental Parameters

Our proposed DBIF-AUNet is implemented using the deep learning framework PyTorch 1.7.0 and Python 3.7.10, and trained using Intel® Xeon® E5-2686 v4 CPU and 2×NVIDIA GeForce RTX 4090 GPU with 24GB memory.

The experiment uniformly adjusts image resolution to 512×512 pixels and uses the SGD optimizer to optimize the training model. The batch size is set to 8, 16, or 24 according to different dataset sizes. The initial learning rate is set to 0.001, and we use a cosine annealing learning rate with restart dynamic update strategy to optimize the training process. The restart step size is set to 20, gamma is set to 0.5, and the training runs for 200 epochs.

#### 4.3 Comparative Experiments

To verify the performance of our optimized model and the quality of the dataset from multiple perspectives, we train other deep learning networks on the constructed dataset and calculate the segmentation results and evaluation metrics of each network. We select U-Net [6], UNet++ [8], ANUNet [24], and Swin-UNet [9] for comparative experiments. In this paper, we choose Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) as core quantitative evaluation metrics, along with commonly used metrics in semantic segmentation models including Accuracy, Precision, Recall, and Specificity. We also visualize the test set prediction results for comparison.

The calculation formulas for IoU and Dice are shown in Equations 25 and 26 respectively:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (24)$$

$$\text{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (25)$$

where  $A \cap B$  represents the intersection area between prediction and ground truth, and  $A \cup B$  represents the union area between prediction and ground truth, calculated as  $A + B - A \cap B$ . IoU focuses more on segmentation accuracy, while Dice focuses more on addressing class imbalance.

We selected currently excellent medical image segmentation networks in the medical deep learning field for horizontal comparison:

Table 1: Comparative experimental results

Model	Year	IoU(%)	Dice(%)	ACC	Pre	Re	Sp
U-Net	2015	71.1	83.1	89.1	83.2	83.0	89.4
UNet++	2018	74.4	85.3	90.4	85.2	85.4	90.9
ANU-Net	2020	75.1	85.8	91.2	85.8	85.8	91.3
Swin-UNet	2021	77.9	87.6	92.7	87.7	87.6	92.1
DBIF-AUNet (Ours)	-	80.1	89.0	94.4	89.1	89.0	94.7

From the data in the table, it can be observed that this dataset achieves good segmentation results on current classical models. Compared to other current medical deep learning models, DBIF-AUNet comprehensively leads in pleural effusion CT image segmentation accuracy with IoU of 80.1

Figure 5 shows the visualization comparison of segmentation results between DBIF-AUNet and the comparison models mentioned above. It can be seen that when facing complex and diverse pleural effusion CT images with unclear edge decomposition for semantic segmentation, our proposed model can more accurately segment complex images and locate boundaries precisely. In contrast, U-Net, UNet++, ANUNet, and Swin-UNet commonly suffer from incorrect segmentation and boundary missing segmentation problems. In summary, the DBIF-AUNet proposed in this paper demonstrates superior segmentation capability on pleural effusion CT images. It can more finely sense changes in image gray values to accurately locate edge boundary segmentation, and shows better accuracy in multi-morphology image recognition without incorrect segmentation for variable effusion types. These results demonstrate its advantages in the field of multi-morphology complex image segmentation with blurred edge boundaries.

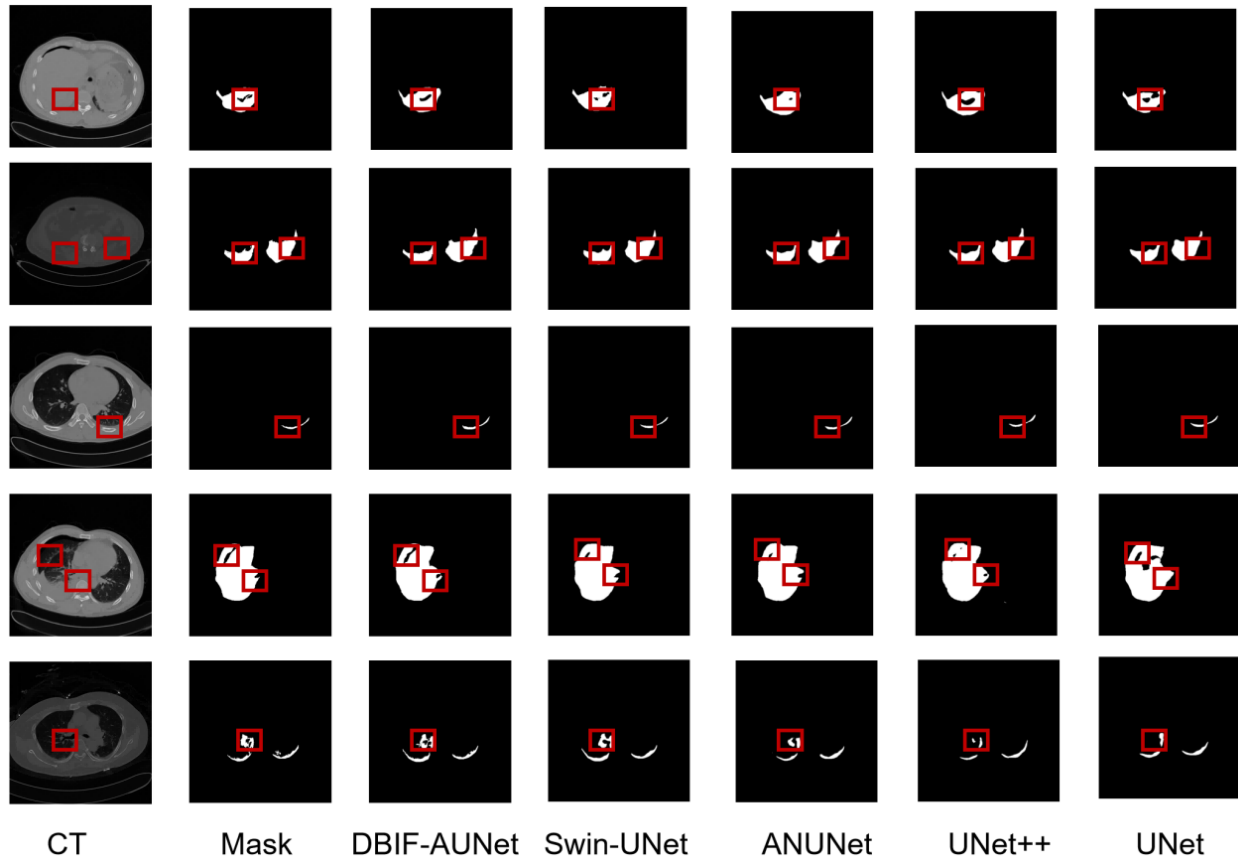


Figure 5: Visual comparison of predictions between DBIF-AUNet and other models

#### 4.4 Ablation Study

To verify the effectiveness of the proposed method with strict variable control, we continue to use the constructed pleural effusion CT image dataset for comparative experiments.

Table 2 shows the results of the ablation study. We conduct ablation experiments by removing or replacing modules, with specific network structure designs as follows:

**DBIF-AUNet:** The complete dual-domain decoupling multi-level interactive attention model proposed in this paper.

**w/o DDFD+BIAF:** This configuration removes the dual-domain decoupling module and multi-level interactive attention fusion module, and deletes the densely nested skip connections designed based on these two modules. It only retains the main network structure and deep supervision module, keeping only the added deep supervision heads after removal.

**w/o DS:** This variant removes the deep supervision mechanism proposed in this paper, retaining other network architecture and the basic deep supervision mechanism of UNet++.

From the information presented in Figure 6 and Table 2, it can be clearly seen that all modules proposed in this paper contribute to improving segmentation accuracy in the network.

Table 2: Ablation study results

Module	IoU(%)	Dice(%)	ACC	Pre	Re	Sp
w/o DDFD+BIAF	74.6	85.4	90.6	85.4	85.4	91.1
w/o DS	79.5	88.6	94.2	88.7	88.5	94.0
DBIF-AUNet (Ours)	80.1	89.0	94.4	89.1	89.0	94.7

Figure 6 shows the visual comparison of model results after ablation experiments:

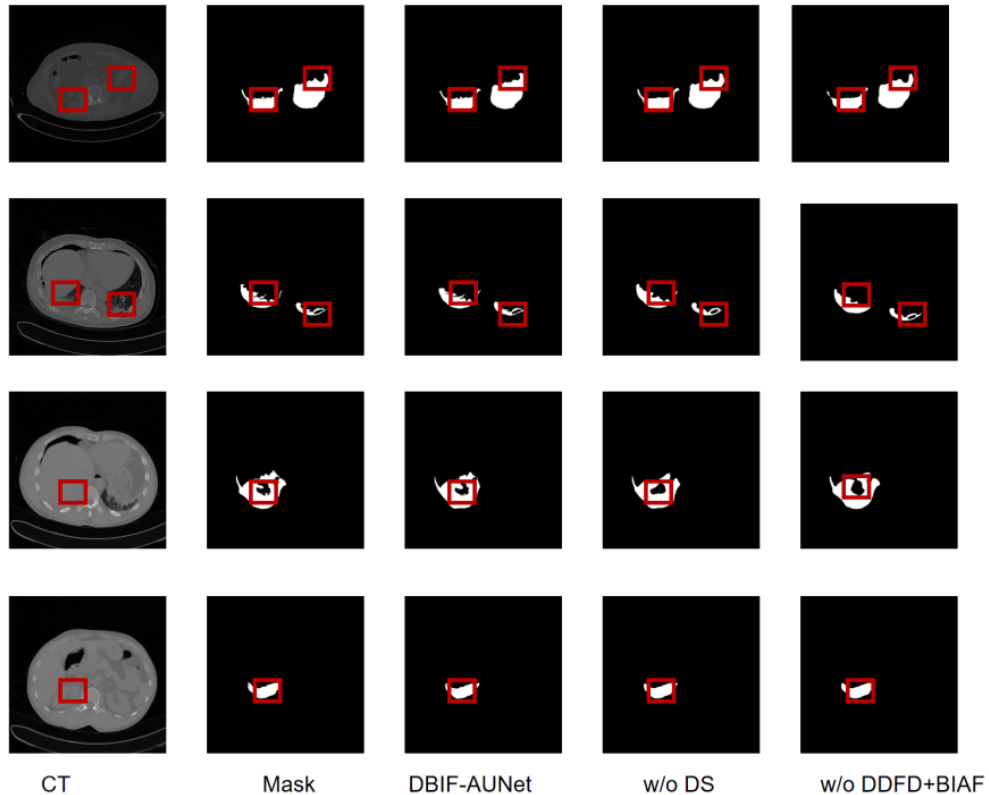


Figure 6: Comparison of ablation experiment results

## 5 Conclusion and Discussion

The DBIF-AUNet proposed in this paper effectively addresses the challenges of diverse effusion types and blurred edges in pleural effusion segmentation through dual-domain decoupling and multi-level interactive attention fusion in a densely nested skip connection network. The model demonstrates excellent performance in edge discrimination and morphological adaptability, particularly showing robust segmentation capability for small amounts of effusion with blurred costophrenic angle features and large-volume circular-like lesions. The nested deep supervision mechanism further improves training stability, accelerates model convergence, and enhances accuracy. When compared with current excellent models in medical image segmentation, our model achieves improvements of at least 2.2

## References

- [1] Kumari, S., & Jaseemudheen, M. M. (2022). Assessment of Pleural Effusion by Means of Imaging Modalities. *Journal of Health and Allied Sciences NU*, 13(03), 301-305.
- [2] Grimberg, A., Shigueoka, D. C., Atallah, A. N., Ajzen, S., & Iared, W. (2010). Diagnostic accuracy of sonography for pleural effusion: systematic review. *Sao Paulo Medical Journal*, 128(2), 90-95.
- [3] Chee, T. M., O'Farrell, H. E., Lima, L. G., Moller, A., Fong, K. M., Yang, I. A., & Bowman, R. V. (2023). Optimal isolation of extracellular vesicles from pleural fluid and profiling of their microRNA cargo. *Journal of Extracellular Biology*, 2(10).
- [4] Han, X., Li, T., Bai, C., & Yang, H. (2024). Integrating prior knowledge into a bibranch pyramid network for medical image segmentation. *Image and Vision Computing*, 143, 104945.
- [5] Li, C., Tan, Y., Chen, W., Luo, X., He, Y., Gao, Y., & Li, F. (2023). Deep Learning-based Semantic Segmentation Methods in Medical Imaging. *Highlights in Science, Engineering and Technology*, 39, 936-942.
- [6] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer.
- [7] Lei, S., Song, J., Wang, T., Wang, F., & Yan, Z. (2024). Attention U-Net based on multi-scale feature extraction and WSDAN data augmentation for video anomaly detection. *Multimedia Systems*, 30(3).
- [8] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6), 1856-1867.
- [9] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv preprint arXiv:2105.05537*.
- [10] Lin, X., Tan, P., Wang, Z., Ma, L., & Li, Y. (2024). DDFL: Dual-Domain Feature Learning for nighttime semantic segmentation. *Displays*, 83, 102685.
- [11] Wei, G., Xu, J., Yan, W., Chong, Q., Xing, H., & Ni, M. (2024). Dual-Domain Fusion Network Based on Wavelet Frequency Decomposition and Fuzzy Spatial Constraint for Remote Sensing Image Segmentation. *Remote Sensing*, 16(19), 3594.
- [12] Zeng, J., Huang, L., Bai, X., & Wang, K. (2024). DBFFT: Adversarial-robust dual-branch frequency domain feature fusion in vision transformers. *Information Fusion*, 108, 102387.
- [13] Hassanin, M., Anwar, S., Radwan, I., Khan, F. S., & Mian, A. (2024). Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108, 102417.
- [14] Yang, X. L., & Jin, Z. Y. (2023). AttGGO-Net: A Semantic Segmentation Method of Lung CT Images with Self and Cross Attention Mechanism. *Journal of Physics: Conference Series*, 2504(1), 012017.
- [15] Chen, B., Huang, X., Liu, Y., Zhang, Z., Lu, G., Zhou, Z., & Pan, J. (2024). Attention-Guided and Noise-Resistant Learning for Robust Medical Image Segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73, 1-13.
- [16] Shen, Y., Guo, J., Liu, Y., Xu, C., Li, Q., & Qi, F. (2025). SMANet: Superpixel-guided multi-scale attention network for medical image segmentation. *Biomedical Signal Processing and Control*, 100, 107062.
- [17] Zhan, B., Song, E., & Liu, H. (2023). FSA-Net: Rethinking the attention mechanisms in medical image segmentation from releasing global suppressed information. *Computers in Biology and Medicine*, 161, 106932.
- [18] Zhao, W., Wei, Q., & Zeng, Z. (2023). A Deeply Supervised Semantic Segmentation Method Based on GAN. *arXiv preprint arXiv:2301.00000*.

- [19] Yuan, D., Xu, Z., Tian, B., Wang, H., Zhan, Y., & Lukasiewicz, T. (2023).  $\mu$ -Net: Medical image segmentation using efficient and effective deep supervision. *Computers in Biology and Medicine*, 160, 106963.
- [20] Dzieniszewska, A., Garbat, P., & Piramidowicz, R. (2025). Deep pixel-wise supervision for skin lesion classification. *Computers in Biology and Medicine*, 193, 110352.
- [21] Liu, X., Zhou, L., Cai, M., Zheng, H., Zeng, S., Wang, X., Wang, Y., & Ding, M. (2025). Hybrid adaptive attention deep supervision-guided U-Net for breast lesion segmentation in ultrasound computed tomography images. *Medical & Biological Engineering & Computing*.
- [22] Maqsood, R., Abid, F., Rasheed, J., Osman, O., & Alsubai, S. (2025). Optimal Res-UNET architecture with deep supervision for tumor segmentation. *Frontiers in Medicine*, 12.
- [23] Li, R., Wang, X., Huang, G., Yang, W., Zhang, K., Gu, X., Tran, S. N., Garg, S., Alty, J., & Bai, Q. (2022). A Comprehensive Review on Deep Supervision: Theories and Applications. *arXiv preprint arXiv:2207.02376*.
- [24] Li, C., Tan, Y., Chen, W., Luo, X., He, Y., Gao, Y., & Li, F. (2020). ANU-Net: Attention-based nested U-Net to exploit full resolution features for medical image segmentation. *Computers & Graphics*, 90, 11-20.