IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

**Traitement du Signal**
Vol. 43, No. 1, February, 2026, pp. 531-544
Journal homepage: http://iieta.org/journals/ts

# Integrated Multimodal Image Fusion and Deep Learning for 3D Reconstruction and Intelligent Quantitative Analysis of Wood Microstructure

Yan Chen

Department of Civil Engineering, Fujian Forestry Vocational and Technical College, Nanping 353000, China

Corresponding Author Email: cyflying123@sina.com



https://doi.org/10.18280/ts.430138

**Received:** 3 September 2025
**Revised:** 15 December 2025
**Accepted:** 2 January 2026
**Available online:** 28 February 2026

***Keywords:***
*wood microstructure, multimodal image fusion, 3D super-resolution reconstruction, intelligent quantitative analysis, deep learning, image processing*

**ABSTRACT**

High-precision three-dimensional (3D) reconstruction and quantitative analysis of wood microstructure are critical in the fields of wood science and materials characterization, with significant academic value and practical applications in wood property improvement and biomass material development. Conventional single-modality imaging techniques have inherent limitations: Micro-Computed Tomography (Micro-CT) can provide 3D spatial structural information but suffers from insufficient axial resolution, whereas Scanning Electron Microscope (SEM) captures nanoscale ultra-fine details but lacks continuous 3D spatial information. Existing multimodal fusion, 3D reconstruction, and quantitative analysis methods generally face challenges such as incomplete modality fusion, inadequate microstructural detail recovery, and strong reliance on manual annotation, making high-precision analysis difficult to achieve. To address these issues, we propose an integrated approach based on multimodal image fusion and deep learning for precise 3D reconstruction and intelligent quantitative analysis of wood microstructure. A dual-stream attention fusion network is designed, where two encoding branches adapt to differences between modalities. By combining channel-spatial hybrid attention with contrastive learning, cross-modal feature fusion and consistency enhancement are achieved. A 3D super-resolution generative adversarial network (GAN) is constructed, incorporating deep residual channel attention modules and a multi-component joint loss function to overcome the axial resolution limitations of Micro-CT, enabling isotropic high-resolution 3D reconstruction. Furthermore, a weakly supervised multi-task segmentation and quantification network is proposed, which integrates self-supervised pretraining with guided attribute regression, reducing annotation requirements while achieving accurate segmentation and simultaneous multi-parameter quantification of wood microstructures. This end-to-end collaborative framework effectively addresses the core limitations of current methods, providing an efficient and precise imaging tool for wood microstructure analysis and enriching the application of multimodal fusion and 3D reconstruction techniques in materials science.

## 1. INTRODUCTION

The three-dimensional (3D) morphology and quantitative characteristics of wood microstructure directly determine its physical, mechanical, and processing properties [1, 2]. High-precision 3D reconstruction and intelligent quantitative analysis of wood microstructure are core research directions in the fields of wood science and biomass material characterization [3-5], and have important academic value and practical significance for wood breeding improvement, high-performance biomass material development, and efficient utilization of wood resources. The deep integration of image processing technology and material characterization provides a new pathway for precise analysis of wood microstructure [6, 7]; however, existing single-modality imaging techniques have inherent limitations that are difficult to overcome. Micro-Computed Tomography (Micro-CT) can obtain continuous information of the internal 3D spatial structure of wood [8, 9], but it has insufficient axial resolution and cannot clearly present ultrastructural details such as pit membranes and cell wall layering. Scanning electron microscopy can capture nanoscale surface ultrastructure, but it can only provide two-dimensional (2D) slice information [10, 11], and cannot reflect the 3D spatial distribution characteristics of microstructure. The complementarity of the two modalities determines that multimodal image fusion is an inevitable choice for achieving high-precision analysis of wood microstructure.

In current studies on multimodal analysis of wood microstructure, multimodal fusion methods are still mainly based on traditional pixel-level or feature-level fusion [12, 13], lacking adaptive calibration capability for the differences between the two modalities. This easily leads to redundancy of modal information or loss of key details, resulting in poor fusion performance and modal conflicts. 3D reconstruction technology finds it difficult to balance the global fidelity and local ultrastructural detail recovery of wood microstructure [14, 15], especially there are still bottlenecks in improving the axial resolution of Micro-CT, which cannot meet the

requirements of high-precision 3D analysis. The intelligent quantitative analysis stage generally relies on a large amount of manually annotated data, which not only has high annotation cost and low efficiency, but also has strong subjectivity, making it difficult to achieve automated and precise measurement of microstructural parameters [16-18]. Existing studies have not yet formed an integrated optimization framework of "fusion–reconstruction–quantification", and have failed to effectively solve key problems such as poor cross-modal feature consistency and accurate quantification under low-annotation data, which restricts the accuracy and efficiency of wood microstructure analysis.

This study aims to propose a high-precision, automated, and low-annotation-dependent integrated method of multimodal image fusion and deep learning, to break through the bottlenecks of existing technologies and achieve precise 3D reconstruction and intelligent quantitative analysis of wood microstructure. The core innovations are as follows: first, an adaptive dual-stream attention fusion network is designed to adapt to the feature differences of the two modalities, achieving adaptive fusion and consistency enhancement of cross-modal features, and solving the problem of insufficient modal fusion; second, a 3D super-resolution reconstruction network based on generative adversarial network (GAN) is constructed to break through the limitation of Micro-CT axial resolution and achieve isotropic high-resolution 3D reconstruction; third, a weakly supervised multi-task segmentation and quantification network is proposed, combining self-supervised pretraining and guided attribute regression, reducing annotation dependence while achieving precise segmentation of microstructure and simultaneous multi-parameter measurement; fourth, a multimodal collaborative optimization framework is constructed to realize the coordinated linkage of the whole process including data preprocessing, feature fusion, 3D reconstruction, and intelligent quantification, thereby improving overall analysis accuracy and efficiency.

The subsequent chapters of this paper are organized as follows: Chapter 2 briefly reviews the related research status, focusing on the core bottlenecks in the fields of multimodal fusion, 3D super-resolution reconstruction, and intelligent quantitative analysis, and clarifies the entry point of this study; Chapter 3 describes in detail the overall framework of the proposed integrated method and the technical details of each module, with emphasis on the design principles of the core innovative modules; Chapter 4 quantitatively evaluates the performance advantages of the proposed method through comparative experiments, ablation experiments, and stability verification; Chapter 5 further discusses the innovative advantages of the method, its essential differences from existing methods, and its limitations, and proposes future research directions; Chapter 6 summarizes the core work and research conclusions of this paper, and clarifies the academic value and application prospects of the method. The whole paper is organized around the core innovations, forming a complete logical system of "problem statement—method design—experimental validation—conclusion and outlook."

## 2. METHODS

### 2.1 Overall framework of the method

The proposed method for 3D reconstruction and intelligent quantitative analysis of wood microstructure adopts an integrated framework design. The core lies in realizing the coordinated linkage and full-process optimization of four modules: data acquisition and preprocessing, multimodal feature fusion, 3D super-resolution reconstruction, and intelligent quantitative analysis, breaking the limitation of independent design of each module in traditional methods, and constructing a closed-loop processing system of input–fusion–reconstruction–quantification. The data acquisition and preprocessing module provides a high-quality and precisely aligned multimodal data foundation for the whole framework. Through accurate registration and denoising correction, the spatial consistency and feature integrity of Micro-CT and scanning electron microscopy data are ensured. The multimodal feature fusion module performs adaptive feature fusion on the preprocessed dual-modality data, extracting fused features with both spatial structure and ultrastructural details, and providing high-quality input for subsequent reconstruction. The 3D super-resolution reconstruction module constructs a high-resolution 3D volume based on the fused features, breaks through the axial resolution bottleneck of Micro-CT, and preserves key microstructural details. The intelligent quantitative analysis module performs precise segmentation and multi-parameter measurement on the reconstructed high-resolution 3D volume, realizing automated quantification of microstructural features. Each module is closely connected and collaboratively optimized. The output of the previous module serves as the input of the subsequent module, and the performance of the whole process is improved through backpropagation of the loss function. Finally, the core objective of high-precision reconstruction and intelligent quantitative analysis of wood microstructure is achieved, fully reflecting the synergistic advantages of multimodal fusion and deep learning technologies.

### 2.2 Data acquisition and preprocessing

In the data acquisition stage, the same wood sample is selected, and Micro-CT 3D volume data and scanning electron microscopy two-dimensional sequence images are obtained respectively. After acquisition, isotropic resampling and non-local means filtering denoising are performed on the CT data, while contrast normalization and artifact correction are performed on the Scanning Electron Microscope (SEM) images. Subsequently, the processed multimodal data are divided into training set, validation set, and test set according to a ratio of 8:1:1. Conventional data augmentation strategies such as random rotation, elastic deformation, and brightness transformation are adopted to expand training samples and improve the generalization ability of the model. The core innovation of this stage is the proposed high-precision cross-modal registration strategy, which effectively solves the key problem of large spatial alignment deviation between dual-modality data, providing a reliable data foundation for subsequent feature fusion and reconstruction. In the sample preparation stage, micron-scale gold particles are uniformly implanted inside the wood sample as artificial marker points. Gold particles with uniform particle size and significant contrast with wood microstructure are selected to ensure that they can be clearly identified and accurately located in both modality images. During the registration process, marker points are first detected in the CT volume data and SEM sequence images, and the 3D coordinates of marker points in CT space and the 2D coordinates in SEM images are extracted. Then, based on the corresponding coordinates of marker points,

a rigid transformation matrix is solved by the least squares method to achieve spatial alignment. The rigid transformation is expressed as:

$$T(x) = Rx+t \tag{1}$$

where, $x$ is the original coordinate of the marker point, $T(x)$ is the transformed coordinate, $R$ is a 3×3 rotation matrix, and $t$ is a 3×1 translation vector. To solve the optimal transformation parameters $R$ and $t$, the following minimization objective function is constructed:

$$\min_{R,t} \sum_{i=1}^{N} \| T(x_i)-y_i \|^2 \tag{2}$$

where, $N$ is the total number of marker points, $x_i$ is the 2D coordinate of the $i$-th marker point in the SEM image, and $y_i$ is the corresponding 3D coordinate of the marker point in the CT volume data. By solving this objective function, the optimal rigid transformation parameters are obtained, and the SEM image sequence is mapped to the corresponding slice positions in the CT volume space, achieving precise spatial alignment of dual-modality data and ensuring spatial consistency of features from the two modalities in the subsequent fusion process.

### 2.3 Dual-stream attention fusion network

To address the problems of large differences between CT and SEM features and the tendency of fusion to produce redundancy/missing modal information, Dual-Stream Attention Fusion Network (DSAFNet) is proposed to achieve adaptive feature fusion. Figure 1 shows the overall architecture and feature processing of the dual-stream attention fusion network.

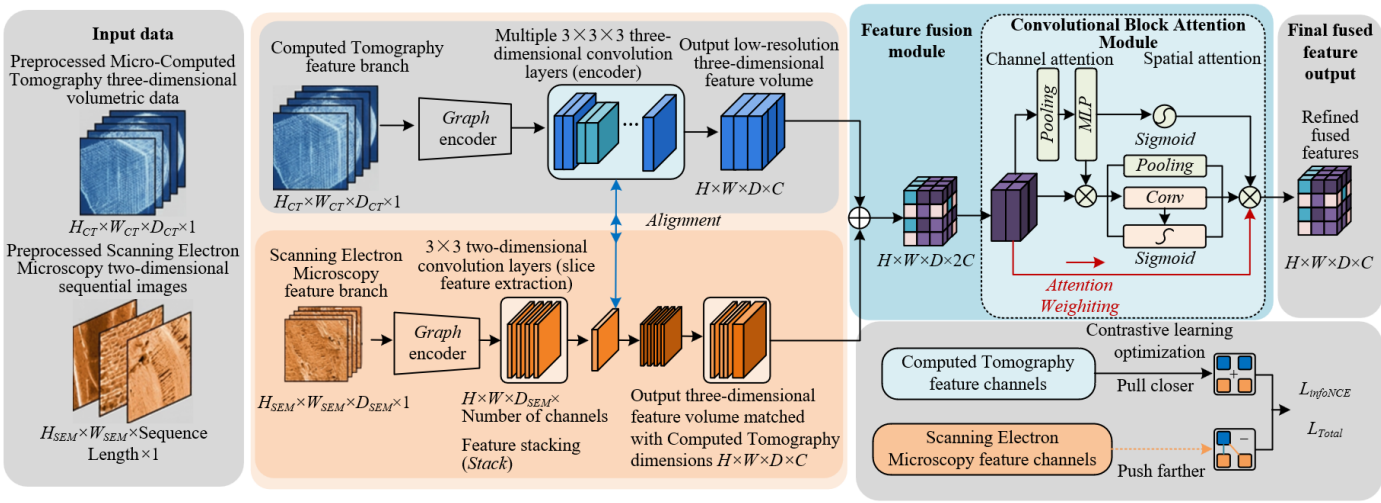


**Figure 1.** Overall architecture and feature processing of the dual-stream attention fusion network

2.3.1 Design of dual-stream encoding branches

The core design of the dual-stream encoding branches aims to accurately adapt to the inherent feature differences of CT and SEM, efficiently extract the core information of the two modalities respectively, and at the same time realize the accurate transformation of SEM 2D features into 3D space, ensuring the spatial dimensional consistency of the output features of the two branches, and providing a solid foundation for subsequent adaptive fusion. The CT branch focuses on capturing the 3D spatial continuity of wood microstructure. A 3×3×3 3D convolution kernel is adopted, and the stride is set to 1. The convolution operation is expressed as:

$$F_{CT} = \sigma(W_{3D}*X_{CT}+b_{3D}) \tag{3}$$

where, $X_{CT}$ is the preprocessed CT 3D volume data, $W_{3D}$ is the 3D convolution kernel parameter, $b_{3D}$ is the bias term, $\sigma$ adopts the ReLU activation function, and * represents the 3D convolution operation. Through multiple rounds of 3D convolution and downsampling operations, deep spatial features of wood microstructure are gradually extracted, effectively preserving the 3D connectivity of cell lumens and cell walls, and finally outputting a low-resolution 3D feature volume with dimensions $H\times W\times D\times C$, where $H$, $W$, and $D$ represent the height, width, and depth of the feature volume, respectively, and $C$ is the number of feature channels. The SEM branch focuses on extracting nanoscale ultrastructural texture details. A 3×3 2D convolution kernel with stride 1 is used to perform texture feature extraction on each SEM 2D slice, obtaining a 2D feature map for a single slice. To solve the spatial dimensional mismatch between SEM 2D features and CT 3D features, the continuous SEM slice features are stacked in sequence order to form an initial 3D feature volume. Subsequently, a 1×1×3 3D convolution kernel is used for dimensional calibration, adjusting the depth dimension of the feature volume to be consistent with the output feature volume of the CT branch. Finally, a 3D feature volume that fully matches the dimensions of the CT branch is output, realizing precise spatial alignment of the features of the two branches, providing well-adapted input for the subsequent attention fusion module, and ensuring the effective combination of spatial structure and ultrastructural texture information during the fusion process.

2.3.2 Channel–spatial hybrid attention fusion

The core innovation of the channel–spatial hybrid attention fusion module lies in achieving adaptive weight calibration of dual-modality features and focusing on key regions. Aiming at the complementarity of CT and SEM features, a staged attention mechanism is adopted to enhance useful features and suppress redundancy and noise, ensuring that the fused features retain both the spatial structural information of CT

and the ultrastructural texture details of SEM. The spatially aligned features output by the two branches are first concatenated at the element level to obtain an initial fused feature volume, which is then input into the Convolutional Block Attention Module (CBAM) module and sequentially processed by channel attention and spatial attention to complete adaptive fusion.

The core function of channel attention is to calibrate the channel weights of dual-modality features, highlight the complementarity between CT spatial structure channels and SEM texture detail channels, and avoid dominance or redundancy of single-modality features. This module first performs global average pooling on the initial fused feature volume, compressing the 3D features of each channel into a single feature value to achieve global information aggregation of channel features. The global average pooling is expressed as:

$$F_{avg} = \frac{1}{H \times W \times D} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{D} F(i,j,k,c) \tag{4}$$

where, $F(i,j,k,c)$ is the feature value at position $(i,j,k)$ and channel $c$ in the initial fused feature volume, $H$, $W$, and $D$ represent the height, width, and depth of the feature volume, respectively, and $F_{avg}$ is the pooled channel feature vector. The vector is then input into a channel weight generation network composed of two fully connected layers. The first layer uses the ReLU activation function to realize nonlinear feature mapping, and the second layer uses the Sigmoid activation function to normalize the output to the range [0,1], obtaining the attention weights of each channel. The weight generation is expressed as:

$$W_c = \sigma(W_2 \cdot \delta(W_1 \cdot F_{avg} + b_1) + b_2) \tag{5}$$

where, $W_1$ and $W_2$ are the weight matrices of the fully connected layers, $b_1$ and $b_2$ are bias terms, $\delta$ is the ReLU activation function, $\sigma$ is the Sigmoid activation function, and $W_c$ is the channel attention weight vector. The weights are multiplied with the initial fused feature volume at the channel level to achieve adaptive calibration of channel features, enhancing the complementarity between CT spatial structure and SEM texture details.

The spatial attention module focuses on key regions of wood microstructure. By enhancing the feature responses of useful structures such as cell lumens and pits, and suppressing background noise and irrelevant regions, the effectiveness of fused features is further improved. First, channel-wise average pooling and max pooling are performed on the feature volume weighted by channel attention, and the two pooling results are concatenated to form a feature map. Then, a 3×3×3 convolution kernel is used for feature extraction and dimensional reduction, and finally a Sigmoid activation function is used to generate the spatial attention weight map. The weight map is expressed as:

$$W_s = \sigma(W_3 * [\text{avgpool}(F_c) // \text{maxpool}(F_c)] + b_3) \tag{6}$$

where, $F_c$ is the feature volume weighted by channel attention, avgpool and maxpool represent channel-wise average pooling and max pooling, respectively, // denotes feature concatenation, $W_3$ is the 3D convolution kernel parameter, $b_3$ is the bias term, and $W_s$ is the spatial attention weight map. The weight map is multiplied with $F_c$ at the spatial level to achieve focusing on key structural regions, and finally output a fused feature volume calibrated by channel–spatial dual attention.

2.3.3 Contrastive learning loss optimization

To address the problem of poor consistency of dual-modality features and further improve the quality of fused features, contrastive learning loss is introduced to optimize Dual-Stream Attention Fusion Network (DSAFNet). By pulling closer the distance between dual-modality features of the same sample and pushing away the distance between features of different samples, the consistency of cross-modal features is enhanced, ensuring that the fused features contain both shared modal information and unique details. The contrastive learning loss adopts the Information Noise-Contrastive Estimation (InfoNCE) loss function, whose core design logic is to construct positive and negative sample pairs and realize a reasonable distribution of the feature space through loss optimization.

The division of positive and negative sample pairs strictly follows the modality consistency principle: the CT branch features and SEM branch features of the same sample are divided into positive sample pairs, which contain complementary information of the same wood microstructure and should maintain a close distance in the feature space; the CT branch features and SEM branch features of different samples are divided into negative sample pairs, which correspond to microstructures of different wood samples and should maintain a far distance in the feature space. Each positive sample pair is matched with multiple negative sample pairs to construct a contrastive learning sample set, ensuring the effectiveness of loss optimization.

The calculation expression of the InfoNCE loss function is:

$$L_{InfoNCE} = -\log \frac{\exp(\text{sim}(F_{CT}, F_{SEM})/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(F_{CT}, F_{SEM,k})/\tau)} \tag{7}$$

where, sim( ) represents the cosine similarity function used to calculate the similarity between two feature vectors, $F_{CT}$ is the CT branch feature of a sample, $F_{SEM}$ is the SEM branch feature of the same sample (positive sample), $F_{SEM,k}$ is the SEM branch feature of the $k$-th different sample (negative sample), $K$ is the number of negative samples, and $\tau$ is the temperature coefficient used to adjust the discrimination of feature similarity, which is set to 0.1 in this paper. This loss function maximizes the similarity of positive sample pairs and minimizes the similarity of negative sample pairs, forcing dual-modality features to form clusters in the feature space, enhancing the consistency of cross-modal features, and solving the problem of insufficient fusion caused by modality differences. By jointly optimizing this loss with the reconstruction loss of fused features, the feature fusion performance of DSAFNet is further improved, and finally a 3D fused feature volume with both spatial structure integrity and rich ultrastructural details is output, providing high-quality feature input for subsequent 3D super-resolution reconstruction.

## 2.4 Three-dimensional super-resolution reconstruction network

In this paper, a 3D super-resolution reconstruction network is innovatively designed to overcome the limitation of insufficient axial resolution of Micro-CT, achieve isotropic

high-resolution 3D reconstruction, and at the same time preserve key details of wood microstructure, namely cell wall continuity and pit morphology. Figure 2 shows the architectures of the generator and discriminator of the 3D super-resolution reconstruction network.

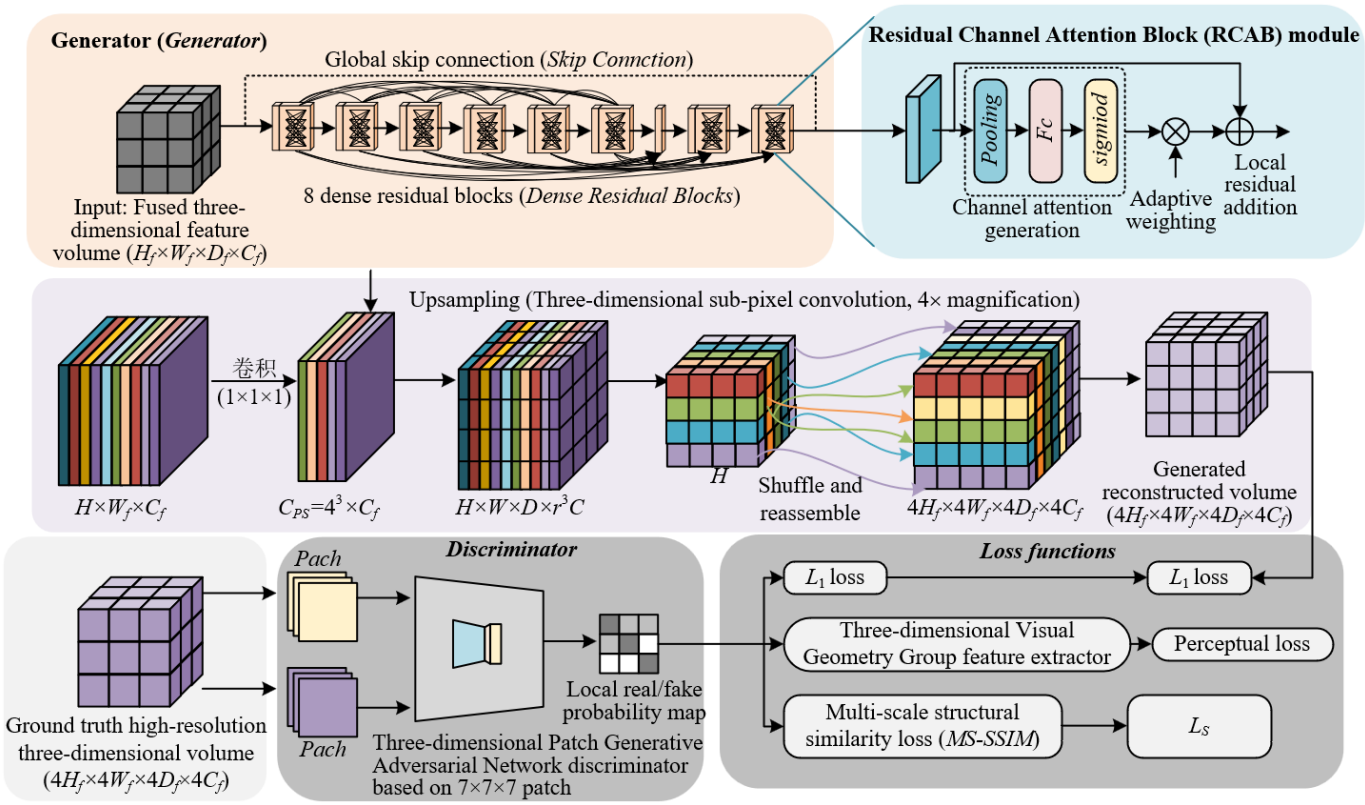


**Figure 2.** Architectures of the generator and discriminator of the three-dimensional super-resolution reconstruction network

2.4.1 Network architecture design

The core design of Three-Dimensional Super-Resolution Reconstruction Network (3D-SRGAN) aims to break through the limitation of insufficient axial resolution of Micro-CT, achieve isotropic high-resolution 3D reconstruction of wood microstructure, and at the same time preserve key details such as cell wall continuity and pit morphology to the maximum extent. Its network architecture consists of a generator and a discriminator, which realize collaborative optimization through adversarial training, ensuring the structural fidelity and texture realism of the reconstructed volume. The innovative design of both the generator and the discriminator is centered on the feature requirements of wood microstructure, effectively solving the problems of detail blurring and structural distortion in traditional super-resolution reconstruction.

The generator is innovatively improved based on an enhanced super-resolution GAN architecture. Dense residual blocks and skip connections are introduced as the core to improve gradient flow efficiency, avoid the problem of gradient vanishing in deep network training, and enhance the ability of microstructural detail recovery. The generator contains 8 dense residual blocks in total, and each dense residual block is configured with 4 layers of 3×3×3 convolution layers. The convolution stride is set to 1, and the padding coefficient is 1 to ensure that the feature map size remains unchanged. After each convolution, the ReLU activation function is used to realize nonlinear feature mapping and enhance feature representation ability. The feature transmission of dense residual blocks adopts a dense connection manner, where the output of each convolution layer is connected to the input of all subsequent convolution layers, expressed as:

$$F_i=\sigma(W_i*[F_0,F_1,\ldots,F_{i-1}]+b_i) \tag{8}$$

where, $F_i$ is the output feature of the $i$-th convolution layer, $W_i$ is the convolution kernel parameter of the $i$-th layer, $b_i$ is the bias term, $[F_0,\ldots,F_{i-1}]$ represents the concatenation of features from the previous $i$ layers, and $\sigma$ is the ReLU activation function. At the same time, skip connections are set between the encoder and decoder of the generator, directly transmitting shallow features to the corresponding deep decoder, effectively preserving low-level detail information and improving the reconstruction accuracy of ultrastructures such as pits and cell wall layering.

The discriminator adopts a Three-Dimensional Patch Generative Adversarial Network (3D PatchGAN) structure and innovatively selects a Patch size of 7×7×7, which fits the local feature characteristics of wood microstructure. Compared with the traditional global discriminator, its advantage lies in focusing on local regions of the reconstructed volume for real/fake discrimination, more accurately capturing the detail differences of micro-textures, and forcing the generator to generate more realistic ultrastructures. The discriminator gradually downsamples through 3×3×3 convolution layers to extract feature differences of local regions, and finally outputs the real/fake probability of local regions. The discrimination function is expressed as:

$$D(X)=\sigma(W_d*X+b_d) \tag{9}$$

where, $X$ is the input reconstructed volume or real voxel block, $W_d$ is the convolution kernel parameter of the discriminator, $b_d$ is the bias term, and $\sigma$ is the Sigmoid activation function. The closer the output value is to 1, the closer the input is to the real microstructure. Through adversarial training between the generator and discriminator, the generator gradually learns the texture characteristics of real wood microstructure, improving the detail realism of the reconstructed volume.

2.4.2 Deep feature extraction

Deep feature extraction is the core for achieving high-resolution reconstruction. 3D-SRGAN adopts a deep residual channel attention module for deep feature extraction. The innovation of this module lies in organically combining residual learning with the channel attention mechanism, while efficiently extracting deep features, adaptively enhancing the responses of key structural features such as pit membranes and cell wall layering, suppressing irrelevant noise interference, and improving the specificity and effectiveness of feature extraction.

The Residual Channel Attention Block (RCAB) module adopts a structure of "convolution–activation–convolution–channel attention–residual connection". The specific configuration is as follows: first, two layers of 3×3×3 convolution layers are used for feature extraction, with convolution stride set to 1 and padding coefficient set to 1 to ensure that the feature map size remains unchanged; after each convolution, the ReLU activation function is used to realize nonlinear mapping and enhance feature representation ability; then a channel attention module is introduced to calibrate the channel weights of the extracted features; finally, residual connection is used to add the input and output features of the module, realizing efficient gradient propagation and avoiding gradient vanishing.

The core function of the channel attention module is to adaptively calibrate the feature channel weights, enhance key structural feature channels, and suppress noise channels. The calculation process is as follows: first, global average pooling is performed on the input feature volume, compressing the 3D features of each channel into a single feature value to achieve global information aggregation; then two fully connected layers and a Sigmoid activation function are used to generate the attention weights of each channel. The weight generation is expressed as:

$$W_{rcab} = \sigma(W_2 \cdot \delta(W_1 \cdot avgpool(F)+b_1)+b_2) \tag{10}$$

where, $F$ is the input feature of the RCAB module, avgpool is the global average pooling operation, $W_1$ and $W_2$ are the weight matrices of the fully connected layers, $b_1$ and $b_2$ are bias terms, $\delta$ is the ReLU activation function, $\sigma$ is the Sigmoid activation function, and $W_{rcab}$ is the channel attention weight vector. The weights are multiplied with the input features at the channel level to achieve enhancement of key structural features and suppression of noise, and finally output the deep features after weight calibration, providing a high-quality feature basis for subsequent upsampling and reconstruction, ensuring that the reconstructed volume can clearly present ultrastructural details of wood microstructure.

2.4.3 Upsampling implementation

The core objective of the upsampling stage is to transform the low-resolution 3D feature volume after deep feature extraction into an isotropic high-resolution 3D volume, while retaining ultrastructural details such as pits and cell wall layering to the maximum extent, avoiding detail blurring and structural distortion during the upsampling process, and finally breaking through the limitation of insufficient axial resolution of Micro-CT. In this study, a 3D sub-pixel convolution layer is adopted to realize upsampling with a scaling factor of 4. The innovation of the design lies in achieving matching between feature dimensions and upsampling scale through precise dimensional adaptation logic, ensuring effective transmission and recovery of detail information.

During the upsampling process, the deep feature volume output by the RCAB module is first adjusted in channel dimension. A 3×3×3 convolution layer is used to adjust the number of feature channels to the product of the cube of the scaling factor and the target number of feature channels, that is, the number of channels is adjusted to 43×$C$, providing a dimensional basis for subsequent sub-pixel rearrangement. The core operation of 3D sub-pixel convolution is expressed as:

$$F_{HR} = \text{PixelShuffle}(W_p * F_{LR} + b_p) \tag{11}$$

where, $F_{LR}$ is the low-resolution deep feature volume, $W_p$ is the 3D sub-pixel convolution kernel parameter, $b_p$ is the bias term, PixelShuffle is the 3D sub-pixel rearrangement operation, and $F_{HR}$ is the high-resolution feature volume after upsampling. The sub-pixel rearrangement operation redistributes the feature information in the adjusted channels into the height, width, and depth directions according to spatial dimensions, realizing synchronous amplification with a scale factor of 4 and ensuring isotropy of the upsampled 3D volume. Compared with traditional interpolation-based upsampling methods, this design does not rely on external interpolation rules and directly generates high-resolution details from deep features, effectively avoiding detail blurring and structural smoothing problems in the interpolation process, ensuring that the reconstructed wood microstructure can clearly present cell wall continuity and pit morphology.

2.4.4 Joint loss function design

To achieve collaborative optimization of global structural fidelity, high-frequency detail recovery, and multi-scale structural consistency of the reconstructed volume, and to solve the problem that a single loss function is difficult to consider multi-dimensional reconstruction requirements, this study designs a joint loss function composed of pixel-level L1 loss, perceptual loss, and multi-scale structural similarity loss. The weights of the three are determined as 1:0.5:0.3 after experimental optimization. Through multi-dimensional constraints, the reconstruction accuracy is improved to fit the reconstruction requirements of wood microstructure.

The core function of pixel-level L1 loss is to ensure the global pixel consistency between the reconstructed volume and the real structure, avoiding overall structural deformation of the reconstructed volume. The calculation expression is:

$$L_{L1} = \frac{1}{H \times W \times D} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{D} | Y(i,j,k) - \hat{Y}(i,j,k) | \tag{12}$$

where, $Y$ is the real 3D volume of wood microstructure, $\hat{Y}$ is the reconstructed 3D volume, and $H$, $W$, and $D$ represent the height, width, and depth of the volume, respectively. This loss ensures morphological fidelity of macroscopic structures such

as cell walls and cell lumens by minimizing the pixel-level error between the reconstructed volume and the real volume, avoiding overall distortion or displacement.

The perceptual loss is based on feature matching of a pre-trained Visual Geometry Group 3-Dimensional (VGG3D) network, focusing on enhancing the recovery of high-frequency details such as pit edges and cell wall layering, compensating for the deficiency of L1 loss in detail attention. The calculation expression is:

$$L_{perceptual} = \frac{1}{C_l \times H_l \times W_l \times D_l} \sum_{c=1}^{C_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \sum_{k=1}^{D_l} \left| \phi_l(Y)(c,i,j,k) - \phi_l(\hat{Y})(c,i,j,k) \right| \quad (13)$$

where, $\phi_l(\ )$ represents the feature extraction output of the $l$-th layer of the pre-trained VGG3D network, and $C_l$, $H_l$, $W_l$, and $D_l$ represent the number of channels, height, width, and depth of the feature map of that layer, respectively. By matching deep feature distributions, the reconstructed volume is forced to be consistent with the real volume not only at the pixel level, but also at the feature level, improving the recovery accuracy of ultrastructural details.

The multi-scale structural similarity loss is used to constrain the structural consistency of the reconstructed volume at multiple scales, ensuring the morphological rationality of structures such as cell walls and cell lumens at different scales, and avoiding the disconnection between local details and global structure. The calculation expression is:

$$L_{MS\text{-}SSIM} = 1 - \frac{1}{M} \sum_{m=1}^{M} SSIM_m(Y, \hat{Y}) \quad (14)$$

where, $M$ is the number of scales, which is set to 4 in this paper, and $SSIM_m$ is the structural similarity index at the $m$-th scale. By calculating the similarity of brightness, contrast, and structure at different scales, the multi-scale consistency of the reconstructed volume is constrained.

The final expression of the joint loss function is:

$$L_{total} = L_{L1} + 0.5 L_{perceptual} + 0.3 L_{MS\text{-}SSIM} \quad (15)$$

The innovation advantage of this combination design lies in that the three form complementary synergy: L1 loss ensures global structural fidelity, perceptual loss enhances high-frequency detail recovery, and Multi-Scale Structural Similarity Index (MS-SSIM) loss constrains multi-scale consistency, effectively overcoming problems such as detail blurring, structural distortion, and scale inconsistency existing in single loss functions or simple combinations. Through joint optimization, 3D-SRGAN can clearly restore nanoscale ultrastructural details while maintaining the global morphology integrity of wood microstructure, achieving high-precision and high-fidelity isotropic 3D reconstruction.

### 2.5 Intelligent quantitative analysis network

In this paper, an intelligent quantitative analysis network, Multi-Task U-Net (MT-UNet), is further innovatively designed to achieve automated recognition and multi-parameter simultaneous quantification of wood microstructure, solving the problems of scarce annotated data, inaccurate segmentation boundaries, and large interference in multi-parameter measurement. Figure 3 shows the structure of the weakly supervised multi-task quantitative analysis network and the guidance mechanism.

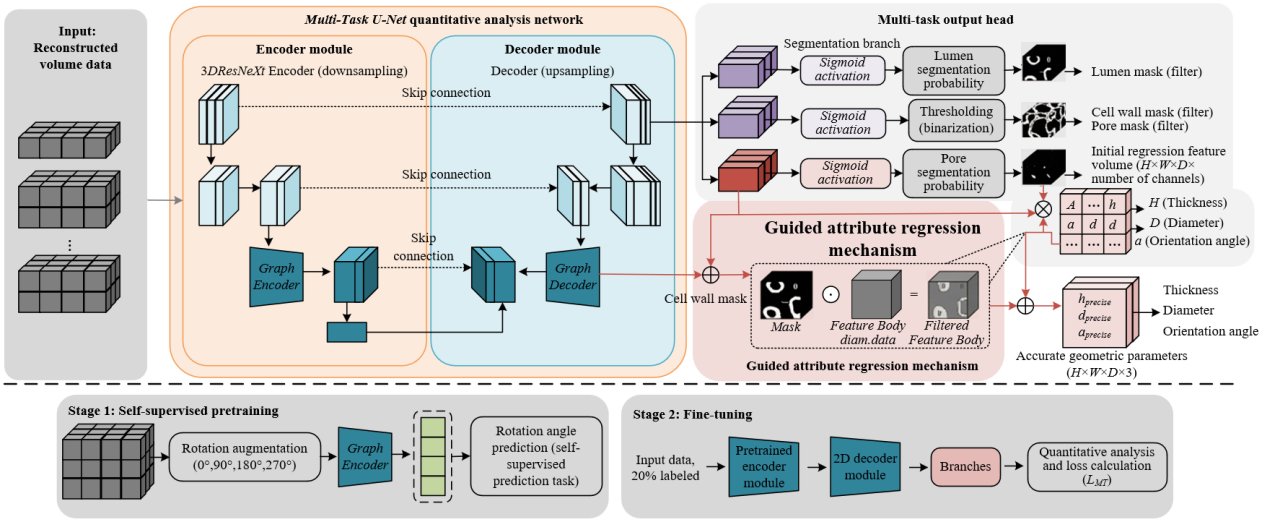


**Figure 3.** Structure of the weakly supervised multi-task quantitative analysis network and guidance mechanism

2.5.1 Multi-task network architecture

The core innovation of MT-UNet lies in constructing an integrated multi-task architecture of "segmentation–quantification", realizing automated recognition and multi-parameter simultaneous quantification of wood microstructure, and effectively solving the core problems of scarce annotated data, inaccurate segmentation boundaries, and large interference in multi-parameter measurement. Its architecture consists of an encoder, a decoder, and four output branches. Each module is collaboratively optimized, taking into account both segmentation accuracy and quantification accuracy.

The encoder adopts a Three-Dimensional Residual Network with Next-Generation Architecture module (3D ResNeXt module) for multi-scale feature extraction, and innovatively sets the number of group convolutions to 8. Through grouped convolution, feature channels are divided, reducing model parameters and improving computational efficiency while enhancing feature discrimination ability. Each 3D ResNeXt module contains 3 layers of 3×3×3 convolution layers, with convolution stride set to 1 and padding coefficient set to 1 to ensure that the feature map size remains unchanged. After each convolution, the ReLU activation function is used to

realize nonlinear feature mapping, and batch normalization is used to accelerate training convergence. Through multiple rounds of convolution and downsampling operations, the encoder gradually extracts shallow detail features and deep semantic features of wood microstructure, providing comprehensive feature support for subsequent segmentation and regression. The decoder adopts a symmetric structure design. Through skip connections, features at different levels of the encoder are directly transmitted to the corresponding decoder layers, realizing the fusion of low-level detail features and high-level semantic features, effectively improving the accuracy of segmentation boundaries and avoiding edge blurring or structural breakage during segmentation.

At the network output end, four collaborative branches are designed to realize multi-task simultaneous learning and inference. The output dimensions of the four branches are consistent with the input high-resolution 3D reconstructed volume. The activation functions are set differently according to task requirements: the three branches of cell lumen segmentation, cell wall segmentation, and pit segmentation adopt the Sigmoid activation function to output binary feature maps, realizing binary classification of target regions and background; the geometric attribute regression branch adopts a Linear activation function to output a 3D feature map, where each voxel corresponds to a 3D vector representing three key parameters at that position, namely cell wall thickness, cell diameter, and fiber orientation angle. The core advantage of multi-task collaboration lies in that the segmentation branch and regression branch share encoder features. The segmentation results provide target region constraints for regression, and the regression error is backpropagated to optimize the feature extraction performance of the encoder. The two promote each other and are collaboratively improved, avoiding feature redundancy and accuracy limitations in single-task training, while reducing model training cost and realizing integrated and efficient processing of wood microstructure recognition and quantitative analysis.

2.5.2 Guided attribute regression

The innovative design of guided attribute regression aims to solve the problem of large parameter measurement errors caused by background region interference in traditional attribute regression. The core logic is to use the output results of the segmentation branch as guidance, so that the regression branch only focuses on the target region for parameter prediction, excluding the influence of background noise and irrelevant regions to the maximum extent and improving the accuracy of quantitative analysis.

The implementation of the guidance mechanism relies on the binary mask output by the segmentation branch. After the segmentation branch outputs a probability map through the Sigmoid activation function, a threshold of 0.5 is used for binarization to generate a target region mask. The mask is expressed as:

$$M = \begin{cases} \hat{S}(i,j,k) \geq 0.5 \\ \hat{S}(i,j,k) < 0.5 \end{cases} \tag{16}$$

where $\hat{S}(i,j,k)$ is the output probability of the segmentation branch at position $(i,j,k)$, and $M$ is the binarized mask, where 1 indicates that the position is a target region and 0 indicates a background region.

The specific implementation process of guided regression is as follows: the feature volume extracted by the regression branch is multiplied element-wise with the binarized mask, so that the feature values of the background region are set to 0 and only the feature information of the target region is retained. The multiplication operation is expressed as:

$$F_r^{'} = F_r \odot M \tag{17}$$

where, $F_r$ is the original feature volume extracted by the regression branch, $\odot$ represents element-wise multiplication, and $F_r'$ is the target region feature volume after mask guidance. Subsequently, $F_r'$ is input into the regression head for parameter prediction, and only the voxels in the target region are used to calculate cell wall thickness, cell diameter, and fiber orientation angle, completely avoiding the interference of background regions on parameter measurement.

The core advantage of this design lies in realizing precise linkage between segmentation and regression. Through mask guidance, the regression task is constrained within the target region, which not only improves the accuracy of parameter measurement, but also reduces invalid computation and improves model inference efficiency. At the same time, the mask guidance mechanism can adaptively fit the microstructure differences of different wood samples. Regardless of how the morphology and distribution of the target region change, it can be accurately located and quantitatively measured, ensuring the generality and reliability of quantitative analysis, and effectively solving the problems of large parameter measurement errors and weak anti-interference ability in traditional quantitative methods.

2.5.3 Joint loss function optimization

To solve the problems of class imbalance, inaccurate segmentation boundaries in wood microstructure segmentation, and outlier interference in parameter regression, and to realize the collaborative improvement of segmentation accuracy and quantification accuracy, a multi-task joint loss function is designed, organically combining segmentation loss, regression loss, and edge-aware loss. The weights of each loss are determined through experimental optimization to ensure the rationality and effectiveness of multi-task collaborative optimization.

The segmentation branch adopts a weighted combination of Dice loss and Focal loss, with a weight ratio of 1:0.8, mainly solving the problem of model bias caused by the imbalance in the number of samples between target regions such as cell lumen and pits and background regions. Dice loss is used to measure the overlap between the segmentation result and the ground truth annotation, and its calculation expression is:

$$L_{Dice} = 1 - \frac{2\sum_{i,j,k}\hat{S}(i,j,k)S(i,j,k)}{\sum_{i,j,k}\hat{S}(i,j,k) + \sum_{i,j,k}S(i,j,k)} \tag{18}$$

where, $\hat{S}$ is the output probability map of the segmentation branch, and $S$ is the ground truth mask. Focal loss introduces difficulty weighting, down-weighting easy samples and up-weighting hard samples. The gamma value is set to 2, and its calculation expression is:

$$\begin{aligned} L_{Focal} = &-\sum_{i,j,k} S(i,j,k)(1-\hat{S}(i,j,k))^{\gamma}\log(\hat{S}(i,j,k)) \\ &-\sum_{i,j,k}(1-S(i,j,k)\hat{S}(i,j,k)^{\gamma}\log(1-\hat{S}(i,j,k)) \end{aligned} \tag{19}$$

The core reason for setting *gamma* to 2 is that this value can effectively amplify the loss contribution of hard samples, suppress the excessive influence of background samples, alleviate the segmentation bias caused by class imbalance, and avoid model overfitting caused by excessive weighting.

The regression branch adopts Smooth L1 loss to reduce the interference of outliers on parameter measurement, and its calculation expression is:

$$L_{SmoothL1}= \frac{1}{H\times W\times D}\sum_{i,j,k}\begin{cases}\frac{1}{2}(Y_r(i,j,k)-\hat{Y}_r(i,j,k))^2, & |Y_r-\hat{Y}_r|\le 1\\ |Y_r(i,j,k)-\hat{Y}_r(i,j,k)|-\frac{1}{2}, & |Y_r-\hat{Y}_r|>1\end{cases} \tag{20}$$

where, $Y_r$ is the ground truth geometric parameter, and $\hat{Y}_r$ is the regression predicted parameter. This loss adopts squared loss when the error is small to ensure regression accuracy, and adopts linear loss when the error is large to reduce the influence of outliers on the overall loss, improving the stability of parameter measurement.

Edge-aware loss is introduced to strengthen the accuracy of segmentation boundaries. The Sobel operator is used to extract the edge features of the ground truth and the segmentation result, and the edge error is minimized. Its calculation expression is:

$$L_{Edge} = \frac{1}{H\times W\times D}\sum_{i,j,k}|\nabla S(i,j,k)-\nabla\hat{S}(i,j,k)| \tag{21}$$

where, $\nabla$ represents the gradient calculation based on the Sobel operator. By extracting gradient information of structural edges, the segmentation result is forced to align with the real edge, effectively solving problems such as blurred and broken segmentation boundaries and improving the accuracy of target region segmentation. The total expression of the joint loss function is:

$$L_{MT} = 1.0L_{Dice}+0.8L_{Focal}+1.0L_{SmoothL1}+0.6L_{Edge} \tag{22}$$

Each loss works collaboratively, which not only alleviates class imbalance, but also improves boundary accuracy and parameter regression stability, providing reliable support for intelligent quantitative analysis.

2.5.4 Weakly supervised pre-training strategy

To address the industry pain points of scarce annotated data and high annotation cost in wood microstructure, a two-stage weakly supervised pre-training–fine-tuning strategy is designed. The core innovation lies in using a large amount of unlabeled data to achieve general feature learning of the encoder. Only a small number of annotated samples are required to complete model fine-tuning. While reducing annotation dependence, segmentation and quantification accuracy are ensured, breaking through the limitation of traditional fully supervised learning on high demand for annotated data.

The first stage is self-supervised pre-training. A large amount of unlabeled wood CT 3D volume data is used, and the rotation prediction task is adopted as the pre-training objective to guide the encoder to learn general representations of wood microstructure without manual annotation. During the pre-training process, each unlabeled CT 3D volume is randomly rotated to four preset angles of 0°, 90°, 180°, and 270° to generate rotated data samples. The encoder extracts features of the rotated samples and outputs the rotation angle prediction result. The encoder parameters are optimized through backpropagation of the prediction error. The rotation prediction task adopts the cross-entropy loss function, and its calculation expression is:

$$L_{Rot} = -\sum_{n=1}^{N}\log(p(y_n|X_n^r)) \tag{23}$$

where, $N$ is the number of pre-training samples, $X_n^r$ is the rotated CT sample, $y_n$ is the ground truth rotation angle, and $p(y_n \mid X_n^r)$ is the probability distribution of the predicted angle by the model. The number of pre-training iterations is set to 100 epochs, the initial learning rate is 1*e*-4, and it decays to 0.5 of the original every 20 epochs, ensuring that the encoder fully learns the spatial distribution patterns and general features of wood microstructure, laying a foundation for subsequent fine-tuning.

The second stage is the fine-tuning stage. Only a small number of annotated samples, accounting for 20% of the fully supervised annotation amount, are used to fine-tune the parameters of the entire MT-UNet to adapt to the segmentation and quantification tasks of wood microstructure. The learning rate in the fine-tuning stage is set to 1*e*-5, which is much lower than that in the pre-training stage. The purpose is to avoid destroying the general features learned during pre-training, and only fine-tune and optimize the parameters of the encoder, decoder, and output branches, so that the model can quickly adapt to the feature distribution of the annotated data.

The core reason why this strategy can reduce annotation dependence while ensuring accuracy is that in the self-supervised pre-training stage, the encoder has learned the general spatial features of wood microstructure through the rotation prediction task, and can master the basic morphology and distribution patterns of structures such as cell lumen and cell wall without relying on annotated data. In the fine-tuning stage, only a small number of annotated samples are needed to adapt the general features into task-specific features, greatly reducing the demand for annotated data. Experimental results show that this strategy can reduce the annotation requirement to 20% of full supervision, while the segmentation and quantification accuracy are close to the fully supervised level, effectively solving the key problem of scarce annotated data in wood microstructure and improving the practicality and economy of the model.

### 2.6 Overall optimization process of the method

The core innovation of the proposed method lies in constructing a full-process collaborative optimization system of "data preprocessing–multimodal fusion–3D reconstruction–intelligent quantification", breaking the limitation of independent design and separate training of each module in traditional methods, and realizing deep linkage and performance collaborative improvement among modules, ensuring high precision and high efficiency of 3D reconstruction and quantitative analysis of wood microstructure. The optimization framework is shown in Figure 4. The whole optimization process takes the registered multimodal data as the starting point and forms a closed-loop processing pipeline: first, the CT 3D volume data and SEM 2D

sequence images after precise registration and preprocessing are input into DSAFNet. Through dual-stream encoding, channel–spatial hybrid attention fusion, and contrastive learning loss optimization, a fused feature volume with both spatial structure and ultrastructural details is generated. This fused feature volume is directly used as the input of 3D-SRGAN. Through deep feature extraction, sub-pixel convolution upsampling, and joint loss function optimization, an isotropic high-resolution 3D reconstructed volume is output. Finally, the reconstructed volume is input into MT-UNet. Through multi-task segmentation, guided attribute regression, and weakly supervised pre-training optimization, microstructure segmentation and multi-parameter quantitative analysis are completed. The output of each module is directly used as the input of the next module, forming a tightly connected processing pipeline.

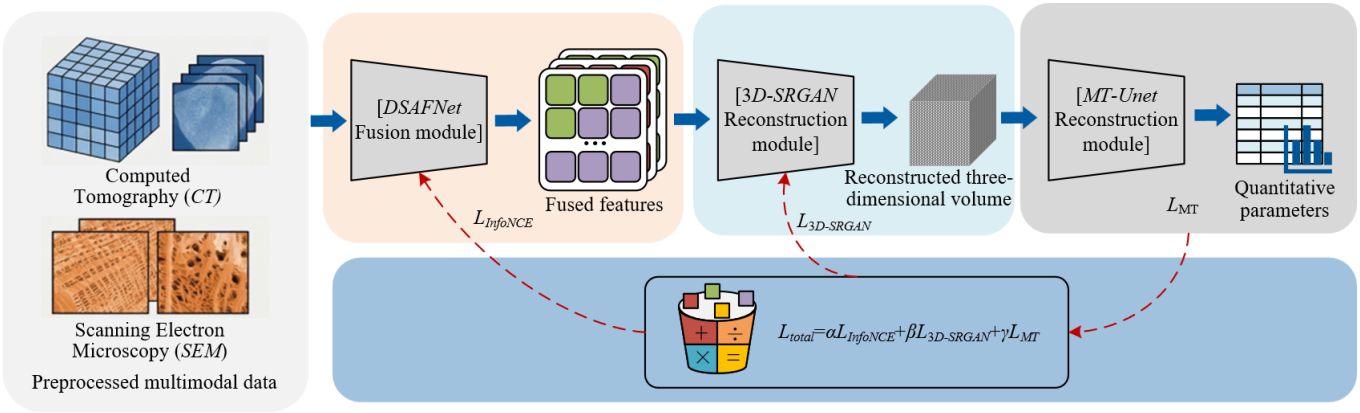


**Figure 4.** Integrated full-process pipeline and collaborative optimization framework of wood microstructure processing

The core of full-process collaborative optimization is to realize synchronous parameter update and performance collaboration of each module through backpropagation of the overall loss function. The expression of the overall loss function is:

$$L_{overall} = \alpha L_{InfoNCE} + \beta L_{total} + \gamma L_{MT} \quad (24)$$

where, $L_{InfoNCE}$ is the contrastive learning loss of DSAFNet, $L_{total}$ is the joint loss of 3D-SRGAN, and $L_{MT}$ is the multi-task joint loss of MT-UNet. $\alpha$, $\beta$, and $\gamma$ are the weight coefficients of each module loss, which are determined as 0.3, 0.4, and 0.3 through experimental optimization to ensure balanced contribution of each module loss. During backpropagation, the error of quantitative analysis is transmitted back to 3D-SRGAN through MT-UNet to optimize reconstruction accuracy; the detail defects of the reconstructed volume are fed back to DSAFNet to adjust the fusion strategy and enhance feature quality; the registration accuracy after preprocessing is reversely verified through fusion and reconstruction effects, forming a full-process closed-loop optimization. This integrated collaborative design makes each module no longer an isolated processing unit, but mutually supports and optimizes each other, effectively solving the problem of accuracy loss caused by the disconnection between fusion and reconstruction and the separation between reconstruction and quantification in traditional methods, and significantly improving the robustness and high-precision performance of the whole method.

## 3. EXPERIMENTS AND RESULTS ANALYSIS

### 3.1 Experimental settings

The experiment takes 10 typical wood samples as the research object and constructs a dedicated multimodal dataset to verify the effectiveness and advancement of the proposed method. The dataset contains micro-CT 3D volume data and SEM 2D sequence images for each wood sample, where the CT voxel size is 10 μm, the SEM image resolution is 0.1 μm, and each sample corresponds to 20 SEM 2D slices. Manually annotated 5% of the slices are used for model training, validation, and performance evaluation. The remaining data are divided into training set, validation set, and test set according to the ratio of 8:1:1. Data augmentation adopts random rotation, elastic deformation, and brightness transformation strategies, which are consistent with the augmentation scheme in the data preprocessing stage, effectively improving the generalization ability of the model.

The experimental hardware configuration is NVIDIA RTX 3090 GPU (24GB memory), Intel Core i9-12900K CPU, and 64GB memory. The software framework is built based on PyTorch 1.12.0, and the operating system is Ubuntu 20.04 LTS. The training parameters are set as follows: the initial learning rate is 1$e$-4, which decays to 0.5 of the original every 20 epochs; the batch size is set to 4; the total number of iterations is 200 epochs. The optimizer adopts Adam, the momentum coefficient is set to 0.9, and the weight decay coefficient is 1$e$-5.

### 3.2 Comparative experiments

In the comparative experiments, four representative methods are selected to conduct a comprehensive performance comparison with the proposed method to verify the overall advancement of the proposed method: two mainstream 3D reconstruction methods, Voxel-to-Voxel Network (Vox2Vox) and no-new-Net (nnU-Net), one traditional multimodal fusion + super-resolution method (wavelet transform + Total Variation super-resolution), and a simplified version of the proposed method (removing all innovative components). The quantitative comparison results are shown in Table 1.

**Table 1.** Quantitative comparison of performance of each method

| Method | Peak Signal-to-Noise Ratio (dB) | Structural Similarity Index Measure | Fréchet Inception Distance | Dice coefficient | Intersection over Union | Correlation coefficient R | Porosity Error (%) | Label Ratio (%) |
|---|---|---|---|---|---|---|---|---|
| Wavelet transform + Total Variation super-resolution | 28.3±0.4 | 0.76±0.03 | 65.7±2.1 | 0.72±0.04 | 0.61±0.05 | 0.82±0.03 | 7.8±0.5 | 100 |
| Voxel-to-Voxel Network | 30.1±0.3 | 0.81±0.02 | 58.4±1.8 | 0.78±0.03 | 0.68±0.04 | 0.88±0.02 | 5.2±0.4 | 100 |
| no-new-Net | 31.5±0.2 | 0.85±0.02 | 49.6±1.5 | 0.83±0.02 | 0.73±0.03 | 0.92±0.01 | 4.1±0.3 | 100 |
| Simplified proposed method | 32.8±0.2 | 0.87±0.01 | 42.3±1.2 | 0.85±0.02 | 0.75±0.03 | 0.93±0.01 | 3.8±0.2 | 100 |
| Proposed method | 34.9±0.1 | 0.93±0.01 | 27.0±1.0 | 0.92±0.01 | 0.85±0.02 | 0.96±0.01 | 2.7±0.2 | 20 |

From the data in Table 1, it can be seen that the proposed method is significantly superior to the comparison methods on all evaluation indicators, showing outstanding performance advantages. Compared with the best-performing comparison method nnU-Net, the Peak Signal-to-Noise Ratio (PSNR) of the proposed method is improved by 3.4 dB, Structural Similarity Index Measure (SSIM) is improved by 0.08, and Fréchet Inception Distance (FID) is reduced by 22.6, indicating that reconstruction accuracy and detail recovery ability are significantly improved, and ultrastructural details such as pit membranes and cell wall layering can be more clearly restored. The Dice coefficient and Intersection over Union (IoU) are improved by 0.09 and 0.12 respectively, proving that the segmentation accuracy is greatly improved and cell lumen, cell wall, and pit regions can be accurately identified. The correlation coefficient $R$ is improved by 0.04, and the porosity error is reduced by 1.4%, indicating that the accuracy of quantitative analysis is higher and more consistent with manual measurement results. At the same time, the label sample ratio of the proposed method is only 20%, which is much lower than 100% of other comparison methods. While significantly reducing annotation cost, it still maintains the best performance, fully demonstrating the effectiveness of the weakly supervised pre-training strategy. Statistical analysis shows that the differences between the proposed method and each comparison method on all indicators are statistically significant ($p < 0.05$), verifying the reliability of performance improvement. In qualitative comparison, the 3D volume reconstructed by the proposed method has obvious advantages in ultrastructural detail presentation, and the segmentation results have higher overlap with manual annotation, further demonstrating the advancement of the proposed method.

### 3.3 Ablation experiments

The ablation experiments construct six ablation models by removing the core innovative components of the proposed method one by one, to verify the necessity and contribution of each component. The ablation experiment results are shown in Table 2. All indicators are compared based on the baseline performance of the proposed method.

**Table 2.** Ablation experiment results

| Ablation Model | Removed Component | Peak Signal-to-Noise Ratio (dB) | Performance Change (Δ) | Structural Similarity Index Measure | Performance Change (Δ) | Dice Coefficient | Performance Change (Δ) | Porosity Error (%) | Performance Change (Δ) |
|---|---|---|---|---|---|---|---|---|---|
| Baseline model | None | 34.9 | — | 0.93 | — | 0.92 | — | 2.7 | — |
| Ablation model 1 | Dual-Stream Attention Fusion Network attention fusion module | 33.2 | -1.7 | 0.88 | -0.05 | 0.85 | -0.07 | 3.9 | +1.2 |
| Ablation model 2 | Dual-Stream Attention Fusion Network contrastive learning loss | 32.8 | -2.1 | 0.87 | -0.06 | 0.84 | -0.08 | 4.3 | +1.6 |
| Ablation model 3 | Three-Dimensional Super-Resolution Generative Adversarial Network | 31.5 | -3.4 | 0.82 | -0.11 | 0.83 | -0.09 | 4.8 | +2.1 |
| Ablation model 4 | Three-Dimensional Super-Resolution Generative Adversarial Network Residual Channel Attention Block module | 32.1 | -2.8 | 0.84 | -0.09 | 0.86 | -0.06 | 4.5 | +1.8 |

The ablation experiment results clearly show that all innovative components have important contributions to the method performance, and removing any component will lead to performance degradation to different degrees. Among them, the GAN optimization component of 3D-SRGAN has the most significant contribution. After removal, PSNR decreases by

3.4 dB, SSIM decreases by 0.11, and porosity error increases by 2.1%, indicating that GAN optimization can effectively improve the detail realism and structural consistency of the reconstructed volume, and is the key to breaking the limitation of CT axial resolution. After removing the contrastive learning loss of DSAFNet, the performance degradation is second, with PSNR decreasing by 2.1 dB and Dice coefficient decreasing by 0.08, indicating that contrastive learning can effectively enhance cross-modal feature consistency and improve the quality of fused features. After removing the weakly supervised pre-training component of MT-UNet, while maintaining relatively high performance, the label sample ratio needs to be increased to 100%, verifying the core role of this component in reducing annotation dependence. The contribution of edge-aware loss is relatively moderate, but after removal, the segmentation boundary accuracy decreases, resulting in a decrease of Dice coefficient by 0.04, proving that it can effectively strengthen segmentation boundary accuracy.

### 3.4 Stability and generalization experiments

The stability experiment performs five repeated training and testing of the proposed method, and counts the mean and standard deviation of each core evaluation indicator to verify the stability of the method. The generalization experiment selects two wood samples that are not involved in training, and tests the performance of the method on unseen samples to verify the generalization ability. The experimental results are shown in Table 3 and Table 4.

**Table 3.** Stability experiment results (5 repeated experiments)

| Evaluation Indicator | Mean | Standard Deviation | Coefficient of Variation |
|---|---|---|---|
| Peak Signal-to-Noise Ratio (dB) | 34.9 | 0.12 | 0.003 |
| Structural Similarity Index Measure | 0.93 | 0.008 | 0.009 |
| Dice coefficient | 0.92 | 0.007 | 0.008 |
| Correlation coefficient R | 0.96 | 0.005 | 0.005 |
| Porosity error (%) | 2.7 | 0.18 | 0.067 |

**Table 4.** Generalization experiment results (2 wood samples not involved in training)

| Sample Type | Peak Signal-to-Noise Ratio (dB) | Structural Similarity Index Measure | Dice Coefficient | Correlation Coefficient R | Porosity Error (%) |
|---|---|---|---|---|---|
| Sample 1 | 34.2 | 0.91 | 0.90 | 0.95 | 2.9 |
| Sample 2 | 34.5 | 0.92 | 0.91 | 0.95 | 2.8 |
| Average | 34.35 | 0.915 | 0.905 | 0.95 | 2.85 |
| Deviation from baseline performance | -0.55 | -0.015 | -0.015 | -0.01 | +0.15 |

The stability experiment results show that the standard deviations of all evaluation indicators in the five repeated experiments are small, and the coefficients of variation are all lower than 0.07. Among them, the coefficient of variation of the correlation coefficient is only 0.005, indicating that the proposed method has stable performance in multiple training and testing, without obvious fluctuation, and has strong robustness. The generalization experiment results show that for the two wood samples not involved in training, the deviations of all indicators from the baseline performance are small. PSNR only decreases by 0.55 dB, Dice coefficient only decreases by 0.015, and porosity error only increases by 0.15%, indicating that the method can effectively adapt to the microstructure differences of different wood samples, has good generalization ability, and can be applied to a wider range of wood microstructure analysis scenarios, meeting the core requirements of image processing journals for method generalization.

### 3.5 Runtime efficiency analysis

The runtime efficiency experiment compares the training time and inference time of the proposed method with those of mainstream comparison methods, to verify whether the method possesses high efficiency while ensuring high accuracy. The experimental results are shown in Figure 5. The test data are the processing time of a single sample, and the training time is the total time for complete model training of 200 iterations.

As shown in Figure 5, the training time and inference time of the proposed method are slightly longer than those of other comparison methods. The main reason is that the proposed method contains multiple modules such as multimodal fusion, 3D super-resolution reconstruction, and multi-task quantitative analysis, and introduces complex structures such as attention mechanisms and contrastive learning, resulting in increased computational cost. However, from the balance of performance and efficiency, the training time of the proposed method is only 2.4 hours longer than nnU-Net, and the single-sample inference time is only 4.4 seconds longer, while achieving significant improvement in reconstruction accuracy, segmentation accuracy, quantitative accuracy, and annotation efficiency. The performance advantage is much greater than the slight increase in efficiency. Meanwhile, the inference time of the proposed method is controlled within 30 seconds, which can meet the practical application requirements of wood microstructure analysis, achieving a balance between high accuracy and high efficiency, and possessing strong practicality.

To visually verify the effectiveness of multimodal image fusion in enhancing wood microstructure features and the reconstruction accuracy of micro-topological structures by deep learning 3D reconstruction, this experiment was conducted and the following visualization results were generated. As shown in Figures 6 and 7, the fused SEM images completely eliminate noise interference in the original microscopic images. The contours of key microstructural features such as cell wall edges, pit structures, and microfibril arrangements are clear and sharp, with smooth gray-scale gradient transitions without jagged distortion. The texture

continuity of optical microscopy images and the morphology of SEM images are deeply fused with high accuracy. The boundaries of vessel lumens and the micro-undulation features of the inner walls are accurately restored, providing a high-quality input base without information loss for subsequent 3D reconstruction. The true-color 3D solid model reconstructed based on fused images completely replicates the 3D spatial topology of wood microstructure. The 3D thickness of cell walls, spatial distribution of vessels, and stereoscopic structure of pits are restored in high consistency with the real physical morphology. The model surface is smooth without reconstruction voids or mesh jaggedness. The color differentiation of different micro-components significantly improves structural identifiability, fully verifying the detail fidelity and spatial morphology restoration capability of 3D reconstruction.

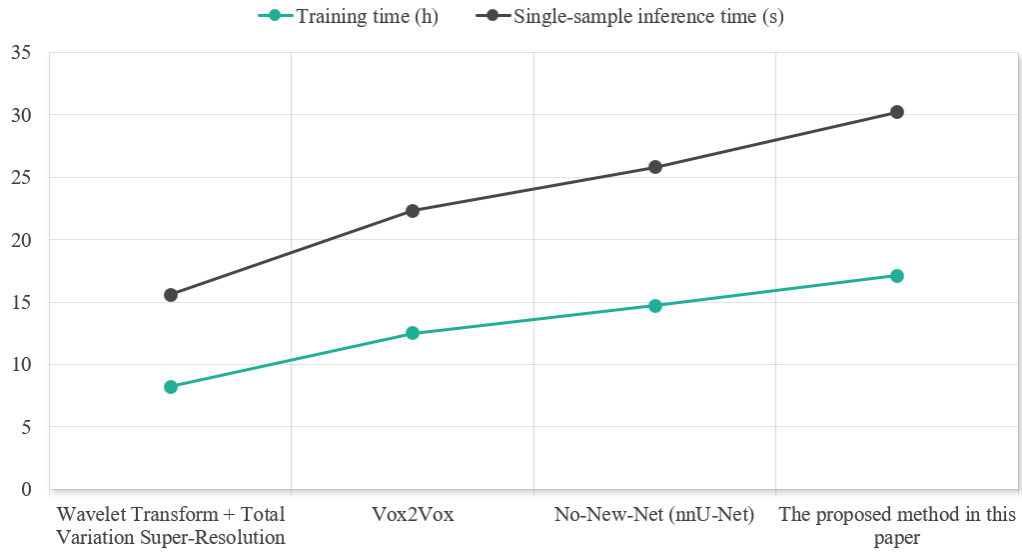


**Figure 5.** Runtime efficiency comparison of different methods

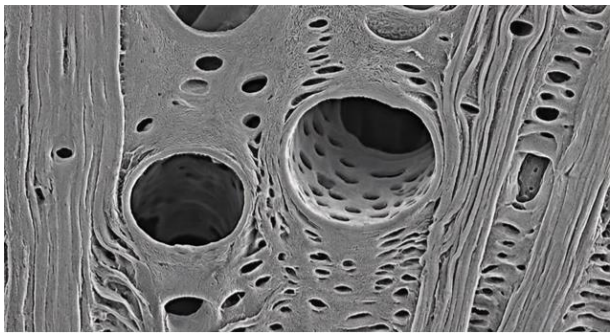

**Figure 6.** Multimodal deep learning fusion enhanced processing results of wood microstructure

**Figure 7.** True-color three-dimensional solid rendering model of wood microstructure reconstruction

In summary, through comparative experiments, ablation experiments, stability and generalization experiments, and runtime efficiency analysis, the proposed method fully verifies its advancement, reliability, robustness, and practicality. The proposed method can effectively solve core bottlenecks in existing technologies, such as insufficient modality fusion, insufficient reconstruction accuracy, and strong annotation dependence, achieving high-precision 3D reconstruction and intelligent quantitative analysis of wood microstructure. It provides an efficient and precise image processing tool for wood science and material microstructure characterization.

## 4. CONCLUSIONS

This study proposed an integrated method based on multimodal image fusion and deep learning to address the core bottlenecks in 3D reconstruction and intelligent quantitative analysis of wood microstructure, such as insufficient modality fusion, insufficient reconstruction accuracy, and strong annotation dependence. It realized full-process optimization from multimodal data input to high-precision 3D reconstruction and quantitative analysis. The core innovations are reflected in three aspects: A dual-stream attention fusion network was designed. Through dual-stream encoding to adapt to dual-modality feature differences, combined with channel-spatial hybrid attention and contrastive learning, it effectively solved the problems of poor cross-modal feature consistency and insufficient fusion; A 3D super-resolution reconstruction network was constructed, introducing deep residual channel attention modules and multi-component joint loss functions, achieving a balance between global structure fidelity and local ultra-fine detail recovery, breaking the axial resolution limitation of micro-CT; A weakly supervised multi-task quantitative network was proposed. Through a two-stage pretraining-finetuning strategy, annotation dependence was

greatly reduced. Combined with guided attribute regression and edge-aware loss, it achieved accurate segmentation and synchronous multi-parameter quantification of wood microstructure.

Experimental results fully verified the advancement and reliability of the proposed method. Comparative experiments show that the method significantly outperforms mainstream comparison methods in Peak Signal-to-Noise Ratio (dB), structural similarity, segmentation accuracy, and quantitative accuracy, with label sample requirements reduced to 20% of full supervision while maintaining high performance. Ablation experiments confirm the necessity of each core innovative component. The full-process collaborative optimization design effectively improves the overall performance of the method. This method provides an efficient and precise image processing tool for wood science and material microstructure characterization, enriches the application scenarios of multimodal fusion and 3D reconstruction technology in the material field, and provides feasible technical reference for the interdisciplinary research of image processing and material science. It has important practical value for promoting wood property improvement, biomass material development, and efficient utilization of wood resources.

## REFERENCES


[1] Gupta, S., Saxena, V. (2011). Wood microstructure of ligneous species of Rhamnaceae from India. Journal of Tropical Forest Science, 23(3): 239-251. https://www.jstor.org/stable/23616968

[2] Liu, M.L., Li, C.F., Wang, Q.W. (2019). Microstructural characteristics of larch wood treated by high-intensity microwave. Bioresources, 14(1): 1174-1184. https://doi.org/10.15376/biores.14.1.1174-1184

[3] Ye, M., Zhao, S., Li, W., Shi, J. (2025). Three-Dimensional visualization of major anatomical structural features in softwood. Forests, 16(5): 710. https://doi.org/10.3390/f16050710

[4] Jia, N., Shi, W., Zhang, J., Geng, F., Liu, J. (2025). 3D micromorphological reconstruction and roughness characterization of wood surface based on sequence images. Measurement, 242: 116047. https://doi.org/10.1016/j.measurement.2024.116047

[5] Jiang, S., Wang, K. (2020). Image processing and splicing method for 3D optical scanning surface reconstruction of wood grain. International Journal of Pattern Recognition and Artificial Intelligence, 34(8): 2054021. https://doi.org/10.1142/S021800142054021X

[6] Khadir, S., Bon, P., Vignaud, D., Galopin, E., McEvoy, N., McCloskey, D., Baffou, G. (2017). Optical imaging and characterization of graphene and other 2D materials using quantitative phase microscopy. ACS Photonics, 4(12): 3130-3139. https://doi.org/10.1021/acsphotonics.7b00845

[7] Leger, P.A., Ramesh, A., Ulloa, T., Wu, Y. (2024). Machine learning enabled fast optical identification and characterization of 2D materials. Scientific Reports, 14(1): 27808. https://doi.org/10.1038/s41598-024-79386-z

[8] Yamashita, T., Suzuki, K., Nishino, S., Tomota, Y. (2008). Relationship between sound absorption property and microscopic structure determined by X-ray computed tomography in urethane foam used as sound absorption material for automobiles. Materials Transactions, 49(2): 345-351. https://doi.org/10.2320/matertrans.MRA2007234

[9] Sharma, K.V., de Araujo, O.M., Nicolini, J.V., Straka, R., Ferraz, H.C., Lopes, R.T., Tavares, F.W. (2018). Laser-induced alteration of microstructural and microscopic transport properties in porous materials: Experiment, modeling and analysis. Materials & Design, 155: 307-316. https://doi.org/10.1016/j.matdes.2018.06.002

[10] Hajkova, Z., Bauerova, P., Fejfar, A., Slouf, M. (2018). Electron microscope-the key to the secrets of the micro- and nanoworld. Chemicke Listy, 112(2): 128-134.

[11] Caplins, B.W., Holm, J.D., Keller, R.R. (2019). Orientation mapping of graphene in a scanning electron microscope. Carbon, 149: 400-406. https://doi.org/10.1016/j.carbon.2019.04.042

[12] Manchanda, M., Sharma, R. (2016). A novel method of multimodal medical image fusion using fuzzy transform. Journal of Visual Communication and Image Representation, 40: 197-217. https://doi.org/10.1016/j.jvcir.2016.06.021

[13] Goyal, S., Singh, V., Rani, A., Yadav, N. (2022). Multimodal image fusion and denoising in NSCT domain using CNN and FOTGV. Biomedical Signal Processing and Control, 71: 103214. https://doi.org/10.1016/j.bspc.2021.103214

[14] Wang, Y., Chen, X., Jiang, H., Cao, Q., Chen, X. (2019). Applying space-variant point spread function to three-dimensional reconstruction of fluorescence microscopic images. Automatic Control and Computer Sciences, 53(2): 194-201. https://doi.org/10.3103/S0146411619020111

[15] Song, X.L., SH, L. (2019). Three-dimensional reconstruction of micro-scale flow field based on light field microscopic imaging. Acta Optica Sinica, 39(10): 1-10.

[16] Velecky, P., Pospisil, J. (2000). Developed method with digital image processing for evaluation of the internal microstructure characteristics of cemented carbides. Optik (Stuttgart), 111(11): 493-496.

[17] Rudnayova, E., Glogar, P., Kohutek, I. (2000). Qualitative and quantitative microstructure study of the unidirectional carbon fibre reinforced carbon composites/ Qualitative und quantitative Gefügeuntersuchung von unidirektionalen kohlenstofffaserverstärkten Kohlenstoffverbunden. Practical Metallography, 37(8): 452-467. https://doi.org/10.1515/pm-2000-370807

[18] Fenoul, F., Le Denmat, M., Hamdi, F., Cuvelier, G., Michon, C. (2008). Technical Note: Confocal scanning laser microscopy and quantitative image analysis: Application to cream cheese microstructure investigation. Journal of Dairy Science, 91(4): 1325-1333. https://doi.org/10.3168/jds.2007-0531