Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

FENet: Frequency-Enhanced Network Based on AFFormer for Wood Surface Defect Detection

Guanghe Cheng[1,3]
Yifei Shao[1,3]
Jinqiang Bai[1,3,*] BAIJQ@SDAS.ORG
Junjie Xia[1,3]
Fengqi Hao[1,2,3]
Yongwei Tang[1,3]
[1]*Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China*
[2]*Faculty of Data Science, City University of Macau, Macau, China*
[3]*Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

The bark is one of the major defects affecting the value of Eucalyptus veneer and must be accurately identified during detection. Currently, for bark defects exhibiting multiple shapes and colors, spatial-domain-based semantic segmentation models often encounter issues with semantic information loss in regions where the color or shape changes, which can lead to incomplete segmentation results. In contrast, utilizing texture features in the frequency domain can reduce the interference caused by variations in color and shape for the model, thereby enabling more effective identification of bark defects. Therefore, in this paper, a Frequency-Enhanced network based on Adaptive Frequency Transformer (AFFormer) is proposed. First, we propose an Inverted Depth-wise Separable Stem to extract more texture information by expanding the number of feature map channels through inverted depth-wise separable convolution. Moreover, we adopt the Rectangular Self-Calibration Module to refine the AFFormer as the backbone network. This enhances the ability to localize bark defects with different shapes and extracts frequency characteristics that are beneficial for semantic segmentation. Finally, the Frequency-Enhanced Channel Attention module enhances the frequency features for semantic segmentation and fuses the spatial-domain feature maps to recover the local details, thereby effectively improving the segmentation accuracy of multi-scale bark defects. Experimental results show that FENet outperforms existing semantic segmentation methods for segmenting bark defects.

Keywords: computer vision, semantic segmentation, wood surface defect detection, supervised learning

1. Introduction

Eucalyptus veneer is widely used in panel manufacturing, but the bark defects on its surface can lead to uneven thickness of the veneer, thus affecting its structural strength and stability. Semantic segmentation can accurately identify the defective areas, so it is widely adopted in the field of defect detection. Chang et al. (2018) studied the impact of different convex optimization weights on the segmentation of wood surface defects. Tabernik et al. (2020) proposed a segmentation detection

© 2025 G. Cheng, Y. Shao, J. Bai, J. Xia, F. Hao & Y. Tang.

method for crack defects in wood, which can achieve good results with a small number of samples. Liu and He (2022) proposed a triple attention semantic segmentation network that uses attention mechanisms to improve the segmentation of small target defects. Ge et al. (2023) introduced a DAF-Net++ model based on an improved U-Net, utilizing VGG16 to enhance image feature extraction capabilities and dense skip connections to integrate semantic information from different levels, thereby improving the segmentation accuracy of wood growth rings. Zhu et al. (2024) proposed a multi-source fusion network based on U-Net that improves the segmentation accuracy of wood broken defects by combining image and depth data. Although the above-mentioned methods have achieved a high accuracy rate in segmenting defects such as live knots on the surface of wood, when dealing with multi-scale defects, there is still the problem of incomplete semantic segmentation. In addition, these methods only focus on the information in the spatial domain and are easily disturbed by the changes in the color or shape of the defects. Moreover, these methods also lack research on the bark, which is a relatively important defect.

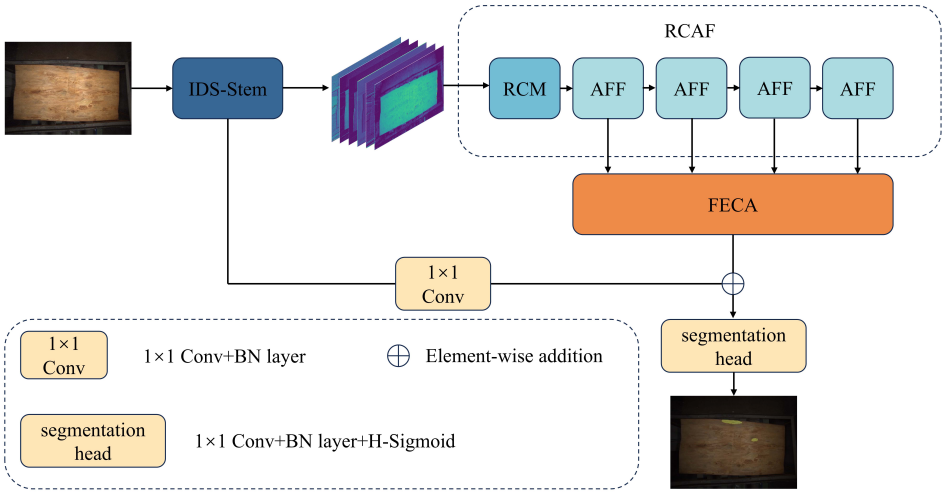


Figure 1: The overall architecture of FENet, which is composed of IDS-Stem, RCAF, FECA, and a segmentation head.

Considering the above analysis, we propose a Frequency-Enhanced Network (FENet) as shown in Figure 1. The texture of the bark varies less compared to the color and shape, so the texture feature is used as an entry point for identifying the bark. Therefore, we propose an Inverted Depth-wise Separable Stem (IDS-Stem) to better highlight the texture information in the feature map. Since the texture features are contained in the high-frequency information of the image, we propose the Rectangular Self-Calibrating Adaptive Frequency Transformer (RCAF) as the backbone network of the FENet to effectively locate bark defects of different shapes and extract the frequency information conducive to segmentation, thereby reducing the interference of color and shape variations on the network. Finally, we propose the Frequency-Enhanced Channel Attention (FECA) module to enhance the frequency information for semantic segmentation, while integrating the spatial-domain information to combine global and local features to improve multi-scale bark defect segmentation accuracy.

2. Methodology

2.1. Inverted Depth-wise Separable Stem

To enable the network to extract richer texture information from the image, we proposed an IDS-Stem. This module is mainly composed of IDS block and 2×2 max pooling, as shown in Figure 2. The IDS block first applies a 1×1 point-wise (PW) convolution to increase the number of channels, thereby enriching the semantic information. Subsequently, a 3×3 DW convolution extracts local features. If down-sampling is necessary, the stride of the 3×3 DW convolution is set to 2 with padding of 1.

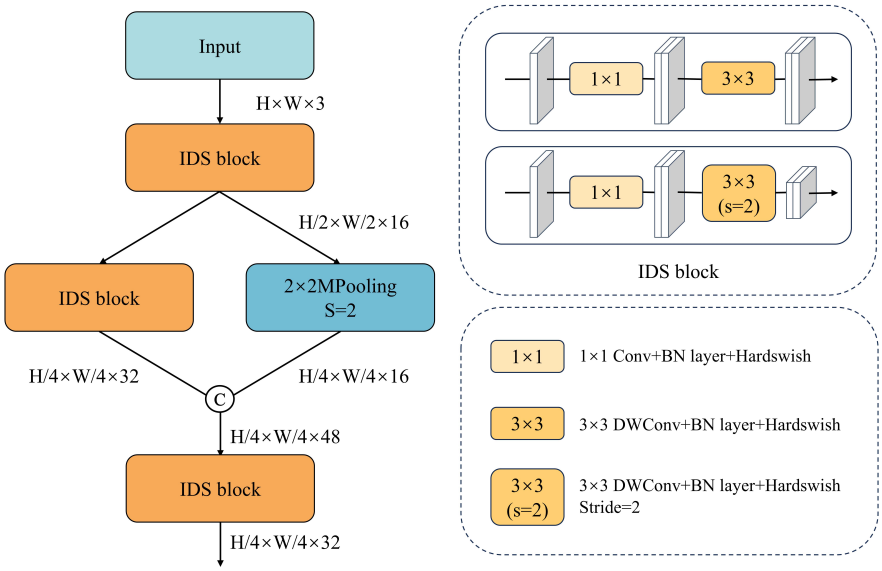


Figure 2: IDS-Stem module structure diagram.

The image is first input into a down-sampling IDS block in the first layer, which preliminarily extracts low-level semantic information while reducing the computational load for subsequent layers. The second layer uses a dual-branch structure: one branch performs down-sampling with an IDS block, while the other uses 2×2 max pooling with a stride of 2. This allows the network to automatically find suitable down-sampling combinations, enhancing the network's ability to capture multi-scale and low-level texture features. The feature maps output by the two branches are concatenated and input into the IDS block of the third layer for feature aggregation, enhancing the network's response to significant features while ignoring redundant information.

2.2. Rectangular Self-Calibration Adaptive Frequency Transformer

Most semantic segmentation networks mainly focus on the information in the spatial domain and utilize relatively little frequency information. However, the texture is a crucial feature for recognizing bark on a veneer, and it is predominantly found in high-frequency information. Moreover, frequency-domain features are relatively insensitive to color and shape changes, which helps to mitigate the impact of these variations on the network. Therefore, we chose to improve the AFFormer (Dong et al., 2023) by integrating a Rectangular Self-Calibration Module (RCM) to propose the

RCAF backbone network, as shown in Figure 1. RCAF's attention can focus on bark defects more effectively, thereby extracting bark texture information more accurately.

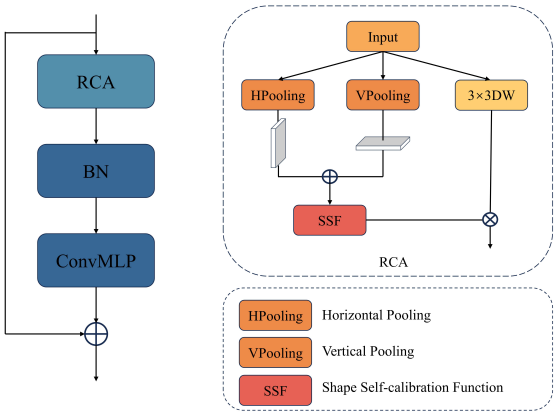


Figure 3: The overall architecture of RCM. The difference between ConvMLP and MLP is that ConvMLP uses convolution to achieve full connection.

The RCAF process is divided into two distinct phases. Due to the diverse sizes and shapes of bark defects, the first phase utilizes the RCM module to augment the network's capability to accurately localize and focus on these specific defect areas. In the second stage, the Adaptive Frequency Filter (AFF) in AFFormer is used to capture frequency information in the feature map that is beneficial for semantic segmentation.

The RCM consists of Rectangular Self-Calibration Attention (RCA) and MLP, as illustrated in Figure 3. RCA uses average pooling to extract global features from both horizontal and vertical directions, generating axial vectors in both directions. These two axial vectors are then summed to form a new feature map, which models the areas of interest in the rectangular regions. Subsequently, a shape self-calibration function is used to refine the areas of interest. This function comprises two large-kernel strip convolutions, calibrating the region shapes from both horizontal and vertical directions, thus enabling adaptation to any shape. This function can be expressed as:

$$F_{ssf}(X) = S(Conv^{l\times1}(ReLU(BN(Conv^{1\times l}(X))))). \tag{1}$$

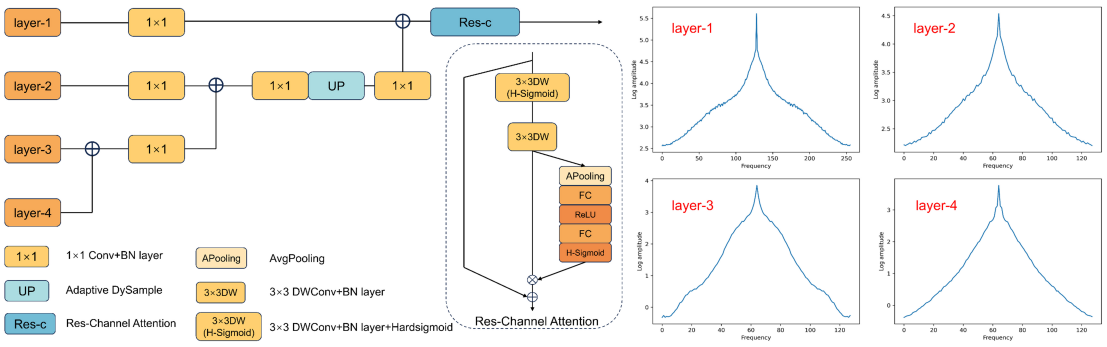
Here, $Conv^{1\times l}$ and $Conv^{l\times1}$ denote the large-kernel strip convolutions in the horizontal and vertical directions, respectively. BN denotes Batch Normalization, $ReLU$ denotes the ReLU activation function, and S denotes the Sigmoid activation function.

Additionally, RCA also includes a 3×3 DW convolution, which is used to further extract local features from the input features. The attention features output by the SSF are weighted with the extracted features through a Hadamard product. Experiments have shown that our proposed RCAF achieves better segmentation results than AFFormer.

2.3. Frequency-Enhanced Channel Attention Module

In the second stage of RCAF, the frequency components and their intensities vary in the frequency information extracted by each layer of AFF, as shown in Figure 4. The first layer does not extract

frequency information and only uses convolution to extract spatial-domain features that contain a lot of local information. The feature maps output from the second and third layers have more concentrated frequency information, contain a higher level of abstraction information, and remove some low and high-frequency noise. The frequency features extracted from the fourth layer are further streamlined to capture mainly global semantic features, but the intensity of high-frequency information is reduced. Therefore, we propose the FECA module as shown in Figure 4. This module is mainly used to effectively integrate the frequency information of each layer and fuse the spatial-domain information to enhance the frequency information that is favorable for semantic segmentation.



(a) FECA

(b) Frequency analysis of feature maps outputs at different layers of AFF

Figure 4: Shows the overall architecture of FECA, where "layer" represents the feature maps output by each layer of AFF.

Our proposed FECA adopts an expansion strategy. First, it uses 1×1 convolution to extend the feature maps of the AFF outputs of each layer in order to match the number of channels in each layer, thus enriching the feature representation. Then it sums the feature map of the fourth layer with that of the third layer and uses 1×1 convolution to integrate the feature information and enhance the intensity of important frequency information. The fused feature maps are then subjected to the same process as the second layer's output and subsequently dynamic up-sampled (Liu et al., 2023) to match the size of the first layer's output feature maps, which recovers the image details and edges more efficiently. The up-sampled feature maps are fused with the feature maps output from the first layer to recover the local detail information lost during the down-sampling process. The resulting output feature maps have a large number of independent channels and rich feature information. However, redundancy may exist in the information of each channel. Therefore, we designed Res-Channel Attention to effectively manage the features across many channels, enhancing those crucial for segmentation while suppressing redundant channels. Res-Channel Attention first uses two 3×3 DW convolutions to independently convolve the features of each channel, thereby enhancing intra-channel feature responses. Then, average pooling is employed to compress the spatial dimensions of each channel to 1. This is followed by two fully connected layers that perform nonlinear transformations on the compressed feature vectors to learn cross-channel dependencies and determine the importance of each channel. Finally, a hard sigmoid function is used to generate the importance weights for each channel, and a Hadamard product is performed with the feature

maps before compression to increase the weights of important channels.

$$F_{feca}(x) = Res(Conv(Up(Conv(Conv(L_4 + L_3) + Conv(L_2)))) + Conv(L_1)) \quad (2)$$

The entire FECA processing procedure is shown in (2). In this context, $Conv$ denotes the 1×1 convolution along with a batch normalization layer, Up represents dynamic up-sampled, and Res signifies Res-Channel Attention. Experimental results have demonstrated that our proposed FECA can significantly improve the performance of the network for semantic segmentation.

3. Experiment

3.1. Experimental Datasets

We collaborated with a wood processing plant to collect a dataset of eucalyptus veneers on-site. The data collection equipment as shown in Figure 5. The collection process was carried out inside the light-blocking box, using LED lights to ensure stable lighting conditions. The eucalyptus veneers measure 61 cm × 127 cm. It is conveyed to the data collection area equipped with optical sensors through a conveyor belt. After the sensors are triggered, the camera takes images of the veneer surface and stores them. Subsequently, images containing bark defects are selected for annotation using Roboflow, resulting in a total of 1503 images of the veneer surface.

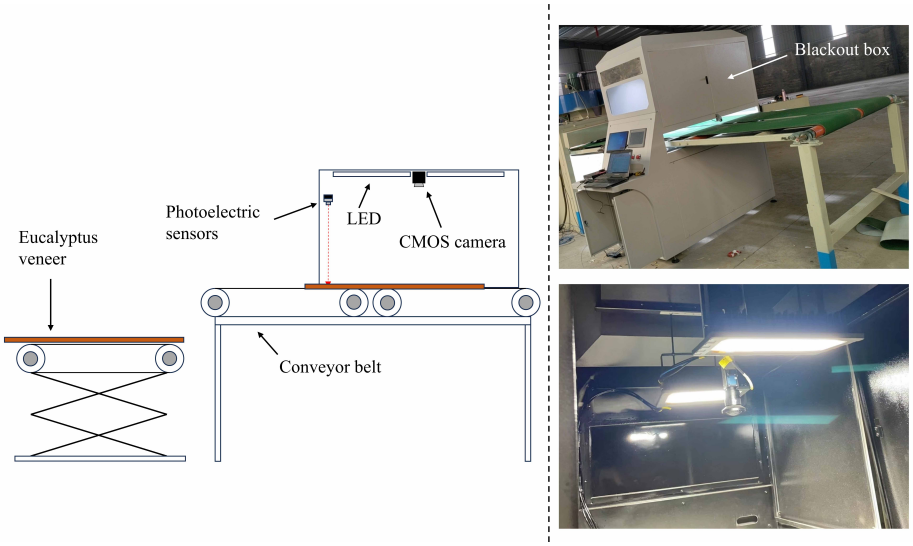


Figure 5: Shows the data collection equipment. The image at the top right is an overall photo of the equipment, and the image at the bottom right is a photo of the inside of the blackout box.

3.2. Implementation Details

All methods involved in the experiments are trained in a Linux environment. Our implementation is based on MMSegmentation and PyTorch and replaces the standard BatchNorm layer with synchronized BatchNorm to facilitate multi-GPU training. For a fair comparison, we use the same data

augmentation strategy as AFFormer and the same CrossEntropy loss function during training. All models were trained on four A100 GPUs with 80K training iterations and a batch size 8. We use AdamW as the optimizer with an initial learning rate of 0.0007, weight decay of 0.01, and polynomial learning rate scheduling strategy coefficient of 1.0. During training, all training images were uniformly resized to 1024×1024.

Table 1: Performance comparison of different segmentation methods.

Methods	Bark.IoU(%)	mPrecision(%)	mRecall(%)	mF1(%)	Bark.F1(%)	Year
HRNet	73.88	95.13	90.11	92.43	84.90	2019
LITv2	75.65	95.52	90.84	93.05	86.13	2022
TopFormer	70.94	94.05	89.19	91.48	82.99	2022
SeaFormer	69.64	94.52	88.06	91.03	82.10	2023
AFFormer	73.58	94.56	90.39	92.37	84.78	2023
CGRSeg	65.60	94.04	85.96	89.59	79.23	2024
FENet(Ours)	77.81	94.68	92.85	93.74	87.52	-

Bark.IoU denotes the IoU of the bark, Bark.F1 denotes the F1 score of the bark.

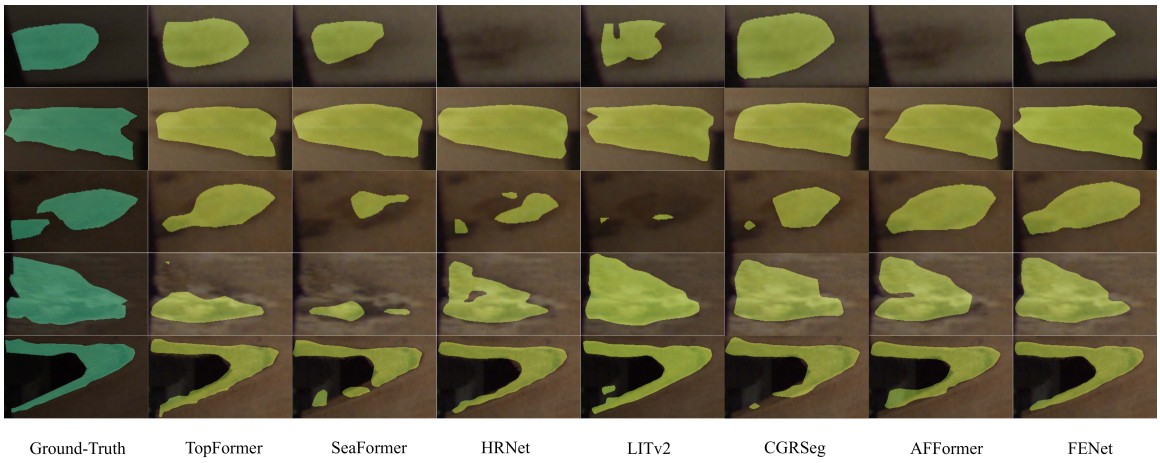


Figure 6: Comparison of segmentation results.

3.3. Comparisons of Different Methods

We compared FENet with several mainstream semantic segmentation methods proposed in recent years, including HRNet (Sun et al., 2019), LITv2 (Pan et al., 2022), TopFormer (Zhang et al., 2022), SeaFormer (Wan et al., 2023), AFFormer, and CGRSeg (Ni et al., 2024). To ensure fairness in the comparison, none of the methods, including ours, used pre-training; instead, they were directly trained on our collected eucalyptus veneer dataset. This approach allows us to assess the network's ability to learn from an unfamiliar dataset. According to the experimental results shown in Table 1, our method significantly outperforms the others in several key metrics, including Bark.IoU, mRecall, mF1, and Bark.F1.

In addition, to better illustrate the segmentation performance of our network, we performed visual analysis, as shown in Figure 6. Segmentation results indicate that FENet, through the integration of spatial and frequency domains, exhibits enhanced robustness, learning capacity, and accuracy in addressing bark defects of various colors, shapes, and sizes.

3.4. Ablation Study

We conducted an ablation study to discuss the effect of each module on the network. In the ablation experiments in this section, the results are shown in Table 2. Compared with the original Stem of AFFormer, our proposed IDS-Stem improves the Bark.IoU by 1.52%. When replacing AFFormer with RCAF as the backbone network, Bark.IoU increases from 73.58% to 74.39%. This indicates that RCAF can focus more on the bark defect region and frequency features are extracted more effectively. The segmentation effect after using the FECA module is significantly better than that of AFFormer, and the Bark.IoU increases from 73.58% to 77.15%.

Table 2: Ablation Study of Each Module.

Setting	Bark.IoU(%)	Bark.F1(%)	mF1(%)
IDS-Stem	75.10	85.78	92.87
RCAF	74.39	85.31	92.64
FECA	77.15	87.10	93.54
IDS-Stem+RCAF	76.63	86.78	93.41
IDS-Stem+RCAF+FECA	77.81	87.52	93.74

To demonstrate the coordination effect between the proposed three modules, we conducted ablation experiments on different combinations of these modules. The experimental results are shown in Table 2. The experimental results show that the segmentation performance is optimal when the three modules work together, which indicates that the collaboration between our proposed modules can effectively improve the network's segmentation performance for bark defects with complex features.

4. Conclusion

This paper describes the FENet method for detecting surface defects in wood. The method utilizes texture information in the frequency domain to effectively address the issue of semantic information loss encountered by existing semantic segmentation models when handling defective regions that exhibit color and shape variations. Additionally, it addresses the problems of incomplete segmentation of large defects and the easy loss of small and edge defects. In the network, we designed IDS-Stem to efficiently capture rich texture details. RCAF was developed from the frequency perspective to enhance the network's ability to locate bark defects and capture frequency information, preventing it from being affected by variations in defect color and shape. We incorporated the FECA module, leveraging both frequency-domain and spatial-domain information to combine global and local features, thereby enhancing segmentation accuracy for bark defects of different sizes. Experimental results show that FENet provides a robust solution for wood surface defect detection with excellent accuracy in dealing with various bark defects.

Acknowledgments

This work was supported by Research and Development of the Composite Sensor for Multiple Soil Parameters and the Intelligent Monitoring System (2024CXGC010905), Research and Application of Key Technologies for Intelligent Management of Agricultural Machinery and Informatization Platform (2023TSGC0587), Research and Development of Key Sensing Technologies for Growth Factors and Vital Signs of Greenhouse Crops (2023TSGC0111) and R&D and Application of Intelligent Central AC and IoT Configuration System for High-end Residences (2024TSGC0603).

References

Zhanyuan Chang, Jun Cao, and Yizhuo Zhang. A novel image segmentation approach for wood plate surface defect classification through convex optimization. *Journal of Forestry Research*, 29 (6):1789–1795, 2018.

Bo Dong, Pichao Wang, and Fan Wang. Head-free lightweight semantic segmentation with linear transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 516–524, 2023.

Zhedong Ge, Ziheng Zhang, Liming Shi, Shuai Liu, Yisheng Gao, Yucheng Zhou, and Qiang Sun. An algorithm based on daf-net++ model for wood annual rings segmentation. *Electronics*, 12 (14):3009, 2023.

Taiheng Liu and Zhaoshui He. Tas 2-net: Triple-attention semantic segmentation network for small surface defect detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6027–6037, 2023.

Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. *arXiv preprint arXiv:2405.06228*, 2024.

Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35:14541–14554, 2022.

Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. *arXiv preprint arXiv:2301.13156*, 2023.

Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022.

Yuhang Zhu, Zhezhuang Xu, Ye Lin, Dan Chen, Zhijie Ai, and Hongchuan Zhang. A multi-source data fusion network for wood surface broken defect segmentation. *Sensors*, 24(5):1635, 2024.