An attention-based wood surface defect detection and classification

Miha Ožbot
Faculty of Electrical Engineering,
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
miha.ozbot@fe.uni-lj.si

Janez Perš
Faculty of Electrical Engineering,
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
janez.pers@fe.uni-lj.si

Abstract

In this paper, we present an approach to image inspection for the detection of wood surface defects based on the attention mechanism. We address the possibility of applying the Swin-Transformer architecture with object detection, localization and classification to the problem of wood surface defect detection, in particular the class of blue stain defects. The visual database of wood surfaces presented for the first time in this study was acquired in collaboration with the local woodworking industry. It consists of 780 images of wood boards of Picea abies and Pinus sylvestris trees and contains 10142 individual bounding box annotations for 18 categories of defects. Our method achieves a mean average precision of 37.1% in a benchmark dataset and 12.7% in our extended defect type dataset.

1. Introduction

In recent decades, a variety of automatic surface inspection (ASI) methods have been proposed to achieve reliable identification of wood surface defects. It has been shown that wood quality grading is a tedious task for humans, resulting in less than optimal detection accuracy [2, 14, 21, 24]. Automated recognition based on computer vision promises to solve this problem while significantly speeding up manufacturing.

Wood surface defects reduce the value of the whole wood plank as they can cause structural weaknesses or limit their use in furniture manufacturing and veneers. In industrial manufacturing, boards are cut according to the segmentation process to remove the defective parts. Therefore, false positive detection of defects is a significant problem, as healthy parts could be discarded, reducing the profit margin. One of the main goals of our algorithm is to minimize the proportion of false-positive detections and thus also the proportion of rejected healthy parts.

A persistently challenging task in wood surface segmentation is the detection of a discoloration called blue stain, which is caused by microscopic fungi, and can occur as a blue, gray, or brown variant. Supposedly, blue stain does not cause structural weakness of the wood but can reduce the esthetic value of the boards used as building or furniture material. Other stains are also possible on the wood surface that are sometimes mistaken for blue stain, such as chemical brown stain caused by drying or lubricating oil from machinery. This defect is mainly visible on the blue channel and therefore requires the use of techniques that include color images. Consequently, this is one of the most

Figure 1. In order of appearance (top left to bottom right), the 18 defect types are: healthy knot (1), healthy knot spike (2), black knot (3), black knot spike (4), fallen knot (5), pith (6), bark pocket (7), rot/decay (8), bark/wain (9), worm holes (10), worm gallery (11), resin pocket (12), fibre damage (13), blue stain (14), brown stain (15), cracks/splits (16), lack of material (17), and resin (18).

challenging tasks in the detection of wood surface defects. Detecting blue stain damage is a major challenge even for a human observer. It does not have a typical shape, it can envelop other defects, it can stretch across a large portion of the observed wood plank, it can be mistaken for other defects, and it requires color information. To overcome these challenges, we use the Swin transformer [18] to address the problem of extracting wood surface features, focusing mainly on the blue stain defect.

2. Related work

An overview of the classically used method for wood surface quality inspection is presented in [14]. In recent years, a variety of deep convolutional neural networks have succeeded in achieving surface defect detection accuracy that meets the industry's 95% accuracy standards. In [24], a faster region-based convolutional network (Faster R-CNN) was used for wood defect detection. A transfer learning method was used for the initial training based on the pretrained neural network models: AlexNet, VGG-16, ResNet, and GoogleLeNet. Transfer learning based on ResNet was also performed in [11]. Transfer learning consistently outperformed training from scratch in all examples. The wood texture was also classified using a different dataset. In another study, the problem of complex non-uniform plank background segmentation was addressed by image binarization and various local thresholding methods [19]. A compact CNN for surface defect detection, classification, and segmentation was presented in [12]. It uses depthwise separable convolution to reduce the number of calculations and weights. A fully convolutional neural network (CNN) was used in [9]. In particular, a polar transform was used to obtain different shapes of wood surface defects. A pair of inception convolutional modules was used to capture features from multiple scales. This study was followed up with a similar 16-layer deep CNN [10]. The authors emphasized the need for an automatic wood feature learning method, as opposed to complex feature extraction. In [26] a CNN feature extractor with an Extreme learning machine (ELM) classifier was used for wood defect box semantic segmentation.

Another approach is to detect anomalies in a predominantly non-anomalous dataset. In [1], a detection of surface anomalies with mixed supervised learning was presented that provided segmentation and classification results. Unfortunately, the wood species that our study focuses on is highly anomalous and such methods would lead to suboptimal detection.

Attention-based transformers, originally intended for language translation and text generation [25], have achieved peak performance in image classification, object detection, and semantic segmentation challenges, but have not yet been applied to wood surface datasets. The attention mechanism main advantage over the Recursive Neural Networks (RNN) used in sequence tasks is that it can be computer in parallel. Transformers are used to solve the problem of varying data sizes and global spatial correlations, which is critical for detecting blue stain defects that can span large portions of a wood plank.

Our work is most similar to the swin transformer [18]. It uses a hierarchical architecture and shifting input windows to solve the problem of multiple object scales. Its successor, the Swin Transformer V2 [17], uses a logarithmic spatial coordinate transformation with relative position distortion to solve the multiple scales problem. The dot product of attention was replaced by a scaled cosine calculation with adaptive scaling to limit the dominance of individual pixels. CoAtNet [4] combines transformers and convolution by adding a kernel to compute self-attention and adding convolution layers before the transformer layer to reduce dimensionality

3. Methods

3.1. Datasets

We introduce a new dataset that consists of 780 RGB color images of wood planks from the spruce (Picea abies) and pine (Pinus Sylvestris) trees. Although this is not a large number of images, these wood types have a large density of defects with an average of 13 defects per image, totaling 10142 defects. Especially prevalent are knot defects with 3527 healthy knot defects and 2632 black knot defects. Conversely, blue stain defects are relatively uncommon with 171 defect occurrences, posing a challenge for deep learning. A heatmap of blue stain defect box annotations is shown in Figure 2. Images of size $W \times H = 4144 \times 384$ pixels were obtained with the use of a SICK Ranger-E55434 camera, which provides a linear scan of the passing plank with a length of 1500 and a width of 130 millimeters. The image acquisition system was shielded from the outside light disturbances and illuminated with a fixed light source but was still subject to random interference from the movement of the wooden plank and the angle of approach. The images were preprocessed to remove the background and roughly realigned to account for the aforementioned plank movement, although some distortion artifacts remained in the data. Defects were labeled with boxed labels in the YOLO format [22] by wood industry experts.

Additionally, a large publicly available benchmark dataset [20] was used to compare results with previous work. This dataset contains 40056 labeled wood surface defects divided into ten defect types, which mostly match ours. Importantly, it contains the blue stain defect with 96 defects, but suffers from the same disproportionality of defects as our dataset with 19457 live knots and 10830 black knots vastly outnumbering other defects. Furthermore, this

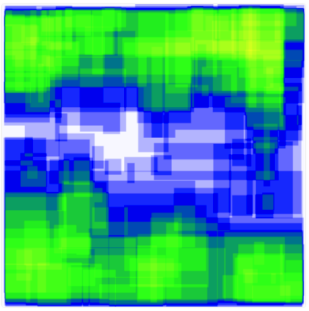

Figure 2. A heatmap of the blue stain defect box annotations from our dataset. The color scale ranges from blue to green to yellow, with transparent blue indicating a single box annotation and yellow indicating a high density of defects. This image suggests that blue stain occurs primarily on the exterior surfaces of the wood planks.

dataset was not acquired in a controlled environment, but rather was in an industrial setting, similar to our dataset. While our dataset consists of images of entire planks, this dataset splits one plank into multiple images.

3.2. Transformers

We experimented with a Transformer feature extraction backbone for wood surface defect detection. The core component of the transformer architecture is a mechanism called attention. It computes meaningful relations between elements in the input data sequence, therefore giving attention to them. An attention function (1) calculates a weighted sum of value vectors V based on a scalar product of key vectors K and query vectors Q

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}\Big)V, \qquad (1)$$

where $\text{Attention}(Q, K, V) \in \mathbb{R}^{M^2 \times d_k}$, $Q \in \mathbb{R}^{M^2 \times d_k}$, $K \in \mathbb{R}^{M^2 \times d_k}$, $V \in \mathbb{R}^{M^2 \times d_k}$ are attention, query, key, and value matrices, respectively; M^2 is the number of patches in a windows; and d_k is the key and query dimensionality.

The vectors V, K and Q are linear transformations of the input signal X with parameters W learned during training. Each Transformer block consists of multiple parallel attention computations collectively called multi-head self-attention (MSA), with their respective keys K, queries Q and values V.

Incidentally, self-attention indicates that all the keys, queries, and value vectors used in the attention calculation come from the same block. This distinction is relevant since the attention mechanism can be used in an encoder-decoder structure where attention and self-attention are both used, i.e., for text translation from one language to another [25]. Multi-head attention is defined as

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_i)W^O \qquad (2)$$

$$\text{head}_i = \text{Attention}(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \qquad (3)$$

where $W_i{}^Q \in \mathbb{R}^{d_k \times d_m}, W_i{}^K \in \mathbb{R}^{d_k \times d_m}, W_i{}^V \in \mathbb{R}^{d_k \times d_m}$ are learnable parameter matrices; i is the head index; and d_m is a model dimension parameter.

The transformer architectures are structurally similar to convolutional neural networks and can benefit from the methods developed for them. Feed-forward layers are used after the MSA structures to compute nonlinear hierarchical features by combining the outputs of all attention heads into one output. This also allows the network to have fixed sizes of the internal layer outputs even when modified during fine-tuning for larger input images. A LayerNorm (LN) is implemented before or after every MSA (implementation differs [17],[18]) to increase computational stability and enable the comparison of feature values between stages, while residual connections solve the problem of exploding/vanishing gradients during training.

Although transformers are effective at storing information in their parameters, they have an inherent disadvantage compared to CNN, namely that they are not invariant to translations in the input data, which is overcome by training transformers with a large number of examples [6]. Training such a model architecture end-to-end is an arduous task, requiring an enormous hand-labelled dataset and GPU processing power. Therefore, a pretrained Transformer model is commonly leveraged as a backbone feature extractor. In the case of object detection, it can be used in conjunction with a classification layer and a region proposal algorithm. The transformer backbone can be fine-tuned or frozen when training the classification head. The latter option is less computationally intensive and usually produces satisfactory results for most tasks.

The aforementioned Swin Transformer [18] is a vision transformer consisting of four stages of multiple successive Swin transformer blocks with patch merging modules between each stage. The number of MSA blocks in each Swin transformer block determines the complexity and size of the model. The smallest version of the pretrained Swin transformer is stylized as Swin-T and is the one we used in this study. It can be used as a backbone structure for a variety of image recognition tasks such as image classification, object detection and classification, and semantic segmentation. The distinctive feature of the Swin Transformer is that it divides the input image into windows and each window into patches, while simultaneously shifting the windows locations in between consecutive MSA layers. This scaling and window crossing enables the detection of differently sized defects and also circumvents the problem of quadratic complexity scaling with image size since this operation scales linearly. The attention mechanism (1) has also been modified with the by adding a relative position bias variable in favor of the input absolute position embedding.

The object classification box annotation average precision (AP) used in the evaluation is calculated as the mean area under the precision-recall curve for every defect class, and for a range of 10 Intersection over Union (IoU) threshold values on the interval [0.5, 0.95] as proposed in the COCO dataset evaluation [16]. Importantly, there is no difference between AP and mean AP (mAP) in the COCO evaluation.

4. Experiments

We perform experiments to compare the attention mechanism with the best convolutional neural networks results for the general wood defect detection task, and we investigate the question of how effective are Transformers at detecting blue stain defects. First, we perform object detection using the classes that match both datasets, i.e., healthy knots, black knots, cracks, resins, and blue stain. In the second test, we use all available defect classes.

The specifications of the hardware and software used for this experimentation are presented in Table 1.

Hardware	Software
GeForce RTX 2060 OC 6GB AMD Ryzen 5 3600X 6-Core 16GB RAM	Python 3.7.5 CUDA 10.1 + cudnn 7.6.3 PyTorch 1.7.1 torchvision 0.8.2 timm 0.3.2

Table 1. Experimental hardware and software environments

Our novel spruce and pine wood surface dataset was converted to a COCO dataset format [16] and divided into subsets of 500, 180, and 100 for training, validation, and testing, respectively. The benchmark dataset was randomly divided into subsets in a 7:2:1 ratio for training, validation, and testing, respectively. The training images were then augmented with 50% probability of horizontal flip, 50% probability of vertical flip, random cropping between 0% and 20% of the image, random shear between -5° and +5° horizontally and -5° and +5° vertically, and random brightness adjustment between -10% and +10%. Using this procedure, we increased both datasets by a factor of three to achieve a better balance of defect types. This is important because this study focuses on an uncommon defect type. Additionally, we performed data augmentation based on the AutoAugment [27] policy during training.

We expected the knot defect types, resin, worm holes, and cracks to attain the highest classification precision, because these defects are the easiest for a human observer to detect and have a distinct outline. In contrast, we expected the stain defects, i.e., blue stain and brown stain, to have low precision for the same reason [23]. Furthermore, some defect classes were extremely rare, i.e., the benchmark overgrowth class and our black knot spike, each having only 7 examples. This small number of examples is obviously not sufficient to train a classification model.

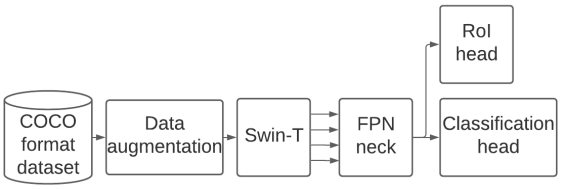

Figure 3. The Swin Transformer backbone architecture with a Feature Pyramid network neck, an object detection head, region proposal algorithm head.

Dataset	Average Recall $\mathrm{AR}^{\mathrm{box}}$
our with 5 classes	0.198
our with 18 classes	0.193
benchmark [20] with 5 classes	0.450
benchmark [20] with 10 classes	0.490

Table 2. Average Recall (AR) comparison of the the Faster R-CNN with Swin-T backbone inference on the test images. The $\mathrm{AR}^{\mathrm{box}}$ denots the box annotation AR for IoU values between 50 and 95.

Dataset	Average Precision $\mathrm{AP}^{\mathrm{box}}$ $\mathrm{AP}^{\mathrm{box}}_{50}$ $\mathrm{AP}^{\mathrm{box}}_{75}$
our with 5 classes	0.133 0.261 0.117
our with 18 classes	0.127 0.292 0.074
benchmark [20] with 5 classes	0.337 0.741 0.260
benchmark [20] with 10 classes	0.371 0.754 0.294

Table 3. Average Precision (AR) comparison of the the Faster R-CNN with Swin-T backbone inference on the test images. The $\mathrm{AP}^{\mathrm{box}}$ denots the box annotation AP for IoU between 50 and 95, $\mathrm{AP}^{\mathrm{box}}_{50}$ denotes the AR for IoU over 0.50, and $\mathrm{AP}^{\mathrm{box}}_{75}$ denotes the AR for IoU over 0.75.

The Swin-T backbone [18] was pretrained on the ImageNet-21k image classification dataset [5]. The two datasets were compared in the inference task with the trained classifier head Faster R-CNN [8].To improve the feature quality, a Feature Pyramid Network (FPN)[15] was used with the hierarchical Swin transformer feature maps from each of the four stages. The Adam optimization technique [13] was used for the optimization of model parameters. It is a stochastic gradient descent method that uses the first and second order moments. The optimizer settings were constant for all experiments, i.e., the learning rate was chosen to be 10^{-4}, the betas were 0.9 and 0.999, and the weight decay was 0.05. All classifiers were trained for 12 epochs.

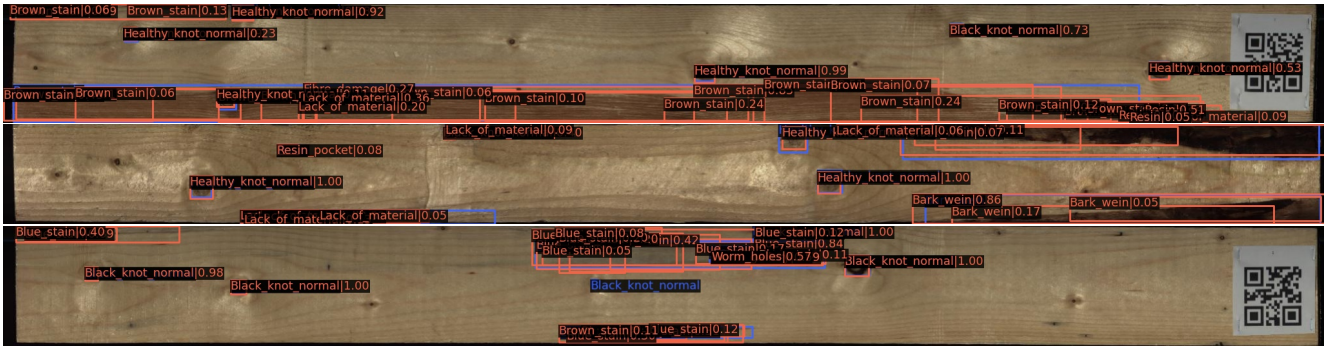

Figure 4. Three examples of object detection inference with the Faster R-CNN with Swin-T backbone of our test dataset with 18 defect categories. The top image shows the detection of the brown stain defect, the middle image shows the detection of the wein defect, and the bottom image show the detection of the blue stain defect. A large number of knot defect were deemed by experts to be too small to be labelled. The stain defect types localization is noticeably more subdivided than the ground truth annotation.

5. Results

First, we trained the Faster R-CNN head with Swin-T backbone for object detection on our dataset and the benchmark dataset separately, using five classes that are present in both datasets. For the second test, we repeated the training of the classifier for object detection on both datasets for all classes. We compared the results of the four detectors to determine how effective the transformer architecture is at detecting the most common wood surface defects and a variety of wood surface defect classes. The results of the two experiments are presented as box annotation Average Recall (AR) in the Table 2 and as mean Average Precision (mAP) or Average Precison (AP) in the Table 3. Examples of inference with the trained object detection model are shown in Figure 4 and Figure 5 for our test dataset and the benchmark test dataset, respectively.

The classifier trained with the benchmark dataset achieved 2-3 times higher AR and AP values than the one trained with our dataset. We attribute this to thorough, consistent, and accurate annotation, that enabled higher localization and classification precision. There is no significant difference in the inference precision of the five common defect classifiers and the classifiers with all defects. This is most probably because the common defect types vastly outnumber all other defect types combined and represent the majority of the $\mathrm{AP}^{\mathrm{box}}$ measure. The recall $\mathrm{AR}^{\mathrm{box}}$ values are consistently higher than the precision $\mathrm{AP}^{\mathrm{box}}$ in all experiments. Higher recall values imply that the ground truth objects were detected, but there are many false positive defections that lower the detection precision.

Interestingly the black normal knot defect precision is lower than expected (25%), even compared with the precision of other defect types. The majority of this discrepancy is the result of false negative defect detection (68%), meaning that the defect was annotated with a box as ground truth but the model did not detect it or detected a smaller area.

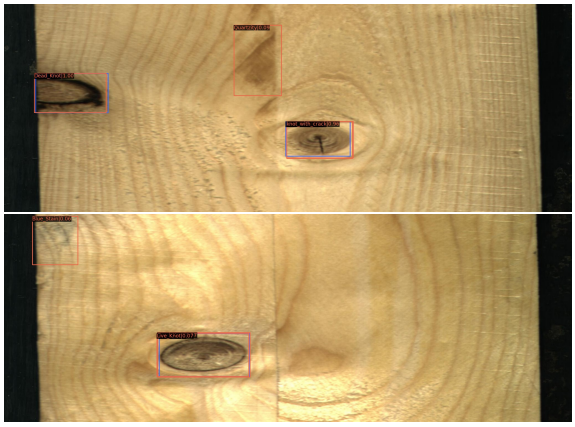

Figure 5. Two examples of object detection inference with the Faster R-CNN with Swin-T backbone of the benchmark test dataset with 10 defect categories. The top image shown a correctly labeled dead knot and a healthy knot with crack and a wrongly classified quartzity defect. The bottom image shown a blue stain defect. Arguably both wrongly classified defects are actually correct, but were not annotated as ground truths. This exemplifies the difficulty of labeling wood defect datasets.

Most surprising is the high results of rare defects such as the rot and decay defect (100%) and the bark pocket defect (45%). Conversely, the cracks and splits defect (11%) and the lack of material defect (15%) are the 4-th and 5-th most common defect types, respectively, but achieved a low AP value. These defect are defined broadly, encompassing defects that could be subdivided into different classes. The worm gallery defect and worm holes defect were frequently mistaken for one another (20%). Similarly, the healthy spike knot was classified for a healthy normal knot (33%) more often than the ground truth (0%).

The detection of blue and brown stain achieved an average precision that is comparable with the other defect classes. This shows the effectiveness of using a transformer architecture to detect stain defect types.

Methods	Type	Number of defects	Accuracy measure
Swin-T, Faster R-CNN (Our)	object detection	18	AP 13%
Swin-T, Faster R-CNN (Our)	object detection	10	AP 37%
ELM, NSST, CNN, SLIC [26]	object detection	N/A	Accuracy 96.72%
LW, ASPP, CNN [12]	object detection	18	N/A
Faster R-CNN, ResNet152 [24]	object detection	5	Accuracy 96.1%
DCNN [10]	object detection	3	Accuracy 99.13%
TL-ResNet34 [7]	image classification	7	Accuracy 98.69%
ResNet18 [11]	image classification	4	Accuracy 99.5%
Mix-FCN [9]	instance segmentation	6	OCA 99.14%
CART, SSIM, Otsu segmentation [3]	instance segmentation	4	Accuracy 94.1%
Local threshold [19]	image segmentation	N/A	Accuracy 92.6%

Overall classification accuracy (OCA), Region-based Convolutional Neural Network (R-CNN), Deep Convolutional Neural Network (DCNN). Extreme learning machine (ELM), Nonsubsampled shearlet transform (NSST), Simple linear iterative clustering (SLIC), Mixed fully convolutional Neural Network (Mix-FCN), Classification And regression tree (CART), Structural similarity (SSIM)

Table 4. Comparison of recent methods used for wood surface defect detection.

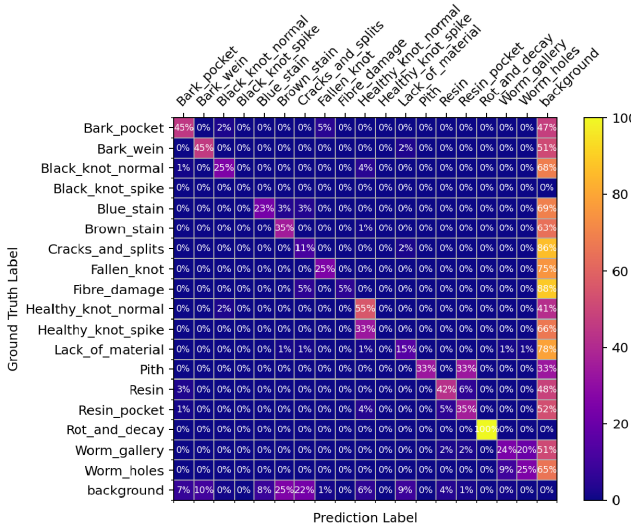

Figure 6. Normalized confusion matrix with ground truth of the AP^{box} object detection inference with the Faster R-CNN with Swin-T backbone of our dataset test images with 18 defect classes.

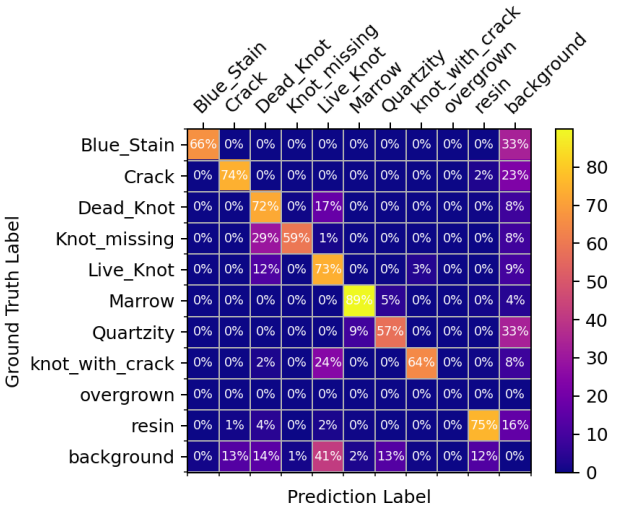

Figure 7. Normalized confusion matrix with ground truth of the AP^{box} object detection inference with the Faster R-CNN with Swin-T backbone on the benchmark dataset test images with 10 defect classes.

Finally, a comparison of the results with our spruce and pine dataset and the bechmark dataset with recent wood surface defect detection methods is presented in Table 4.

6. Conclusion

In this study, we demonstrated that the attention based transformer architecture, specifically the Swin-T transformer, can be used to extract wood surface defect features for use in a object detection and classification task. It can be used to detect stain type defect as it is capable of extracting defect features spanning the length of the whole image and it can detect classes of varying sizes.

We trained a Faster R-CNN classifier with a Swin-T backbone to extract image features on an original wood surface dataset that was first presented in this study, and a benchmark wood surface dataset. Basic data augmentation was performer to enlarge both datasets, i.e., vertical and horizontal flip, random image crop, vertical and horizontal sheer, and brightness change. We compared training classifiers on the five most common defect classes and all available classes from both datasets.

The blue stain defects as well as the brown stain defects are hard to detect and differentiate even for humans, since it is hard to annotate their fuzzy borders with box annotations.

This subjective bias is present for all defect classes, including the knot defects, i.e, small knots might not be labeled, while black knots might be mislabeled as healthy knots and vice versa. A large dataset of defects is required to mitigate the effects of this inherit bias, but the current results are promising.

7. References

[1] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 8 2021.

[2] Ibrahim Cetiner, Ahmet Ali Var, and Halit Cetiner. Classification of knot defect types using wavelets and knn. *Elektronika ir Elektrotechnika*, 22:67–72, 12 2016.

[3] 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*, 29:1789–1795, 11 2018.

[4] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. 6 2021.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 10 2020.

[7] Mingyu Gao, Jianfeng Chen, Hongbo Mu, and Dawei Qi. A transfer residual neural network based on resnet-34 for detection of wood knot defects. *Forests 2021, Vol. 12, Page 212*, 12:212, 2 2021.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Ting He, Ying Liu, Chengyi Xu, Xiaolin Zhou, Zhongkang Hu, and Jianan Fan. A fully convolutional neural network for wood defect location and identification. *IEEE Access*, 7:123453–123462, 2019.

[10] Ting He, Ying Liu, Yabin Yu, Qian Zhao, and Zhongkang Hu. Application of deep convolutional neural network on feature extraction and detection of wood defects. *Measurement: Journal of the International Measurement Confederation*, 152, 2 2020.

[11] Junfeng Hu, Wenlong Song, Wei Zhang, Yafeng Zhao, and Alper Yilmaz. Deep learning for use in lumber classification tasks. *Wood Science and Technology*, 53:505–517, 3 2019.

[12] Yibin Huang, Congying Qiu, Xiaonan Wang, Shijun Wang, and Kui Yuan. A compact convolutional neural network for surface defect inspection. *Sensors 2020, Vol. 20, Page 1974*, 20:1974, 4 2020.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.

[14] Martin Kryl, Lukas Danys, Rene Jaros, Radek Martinek, Pavel Kodytek, and Petr Bilik. Wood recognition and quality imaging inspection systems, 2020.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. 12 2016.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. 5 2014.

[17] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. 11 2021.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 3 2021.

[19] Wei Luo and Liping Sun. An improved binarization algorithm of wood image defect segmentation based on non-uniform background. *Journal of Forestry Research*, 30:1527–1533, 8 2019.

[20] Kodytek Pavel, Bodzas Alexandra, and Bilik Petr. Supporting data for deep learning and machine vision based approaches for automated wood defect detection and quality control., Apr 2021.

[21] Vincenzo Piuri and Fabio Scotti. Design of an automatic wood types classification system by using fluorescence spectra. *undefined*, 40:358–366, 5 2010.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[23] Ruoxu Ren, Terence Hung, and Kay Chen Tan. A generic deep-learning-based approach for automated surface inspection. *IEEE Transactions on Cybernetics*, 48:929–940, 3 2018.

[24] Augustas Urbonas, Vidas Raudonis, Rytis Maskeliunas, and Robertas Damaševičius. Automated identification of wood veneer surface defects using faster region-based convolutional neural network with data augmentation and transfer learning. *Applied Sciences (Switzerland)*, 9, 11 2019.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 2017-December, 2017.

[26] Yutu Yang, Xiaolin Zhou, Ying Liu, Zhongkang Hu, and Fenglong Ding. Wood defect detection based on depth extreme learning machine. *Applied Sciences (Switzerland)*, 10:1–14, 11 2020.

[27] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning data augmentation strategies for object detection. 6 2019.