ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-M-2-2025
30th CIPA Symposium “Heritage Conservation from Bits:
From Digital Documentation to Data-driven Heritage Conservation”, 25–29 August 2025, Seoul, Republic of Korea

An End-to-End AI Pipeline for Wood Knot Detection to Enhance Structural Assessment in Historic Timber Structures

Junquan Pan [1], Maria Chizhova [1], Frank Ebener [2], Thomas Luhmann [3] , Christian Ledig [4], Ferdinand Maiwald [5], Thomas Eißing [1]

[1] KDWT, Otto-Friedrich University of Bamberg, Germany – (junquan.pan, maria.chizhova, thomas.eissing)@uni-bamberg.de
[2] Independent expert, Bamberg, Germany – frank.ebner92@web.de
[3] IAPG, Jade University of Applied Sciences, Germany – luhmann@jade-hs.de
[4] xAILab Bamberg, Otto-Friedrich University of Bamberg, Germany – christian.ledig@uni-bamberg.de
[5] Chair of Optical 3D-Metrology, Dresden University of Technology, Germany – ferdinand.maiwald@tu-dresden.de

Keywords: Photogrammetry, Deep Learning, Historic Timber Structures, Wood Knot Detection, Structural Assessment.

Abstract

The accurate detection and assessment of wood surface defects in historic timber structures, particularly knots, is essential for effective conservation and strengthening planning. However, the application of automated visual grading methods to aged timber remains underexplored due to the irregular texture, weathering and lack of relevant datasets. In this study, we propose an end-to-end deep learning-based pipeline that integrates wood surface segmentation, perspective correction, and knot detection to estimate structural grading factors. A dedicated raw data collection of over 10,000 high-resolution images of historic timber surfaces was compiled using both DSLR cameras and mobile devices, resulting in multiple datasets with approximately 3,000 annotated samples. Three model families, YOLO, Detectron2 and DeepLabV3, were evaluated under different experimental setups. Beyond model benchmarking, we further compared the AI-derived results with expert manual measurements. The model for segmentation of timber surface achieved a mean IoU of over 0.85 and the model for detection of historical wood knots reached F1 scores of up to 0.9. The structural assessment factors estimated by the AI pipeline achieved a Pearson correlation coefficient of 0.641 compared to manual measurements, indicating a moderate level of consistency in knot factor estimation. This research highlights the potential of vision-based AI systems in supporting structural diagnosis and conservation of heritage timber elements.

1. Introduction

1.1 Research background

Over time, various environmental and mechanical factors affect the stability and integrity of historic timber structures. Despite their cultural and architectural significance, these structures are often undervalued due to the lack of accurate assessments of their mechanical strength and structural parameters. Traditional inspection methods, which rely on manual measurements, are time-consuming, prone to human error, and often inadequate for challenging conditions, such as poor lighting or inaccessible areas.

A detailed evaluation of existing timber structures offers the potential to identify and utilise previously unrecognised structural reserves. This, in turn, enables a more precise planning of reinforcement measures during refurbishment projects and may lead to a reduction in the extent of required interventions. For installed timber, strength grading can be performed based on visually detectable characteristics, in accordance with the German standard DIN 4074-1 and DIN EN 14081-1. Among the listed criteria, knots are particularly crucial for determining the strength (Görlacher, 1999). Recent research (Zhang, 2024) investigates the effect of knots on the mechanical properties of Chinese fir using a three-point bending test and X-ray computed tomography, revealing that knot size and position significantly influence strain distribution and mechanical behaviour.

In practice, the visual grading of in-situ timber beams is rarely applied, primarily for two reasons. Firstly, the structural engineer responsible requires extensive additional knowledge of timber as a material. Secondly, the documentation of the grading process is time-consuming and lacks clearly defined regulatory standards. Figure 1, taken from DIN 4074-1, illustrates the method by which knot characteristics in squared timber sections are evaluated. The knot factor A represents the largest ratio between the shortest diameter of a visible knot and the cross-sectional dimensions (width or height) of the timber, based on measurements from all four sides, which can be calculated using the following equation:

$$A = max.\left(\frac{d_1}{b};\frac{d_2}{h};\frac{d_3}{b};\frac{d_4}{h}\right)$$

where d_1, d_2, d_3, d_4 denote the shortest measured diameters of the visible knots, while b stands for the width and h for the height of the timber.

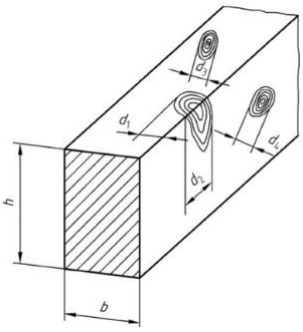


Figure 1. The determination of the largest single branch by the dimension of knot (DIN, 2003).

1.2 Relevance and Related Works

Automatic detection of wood surface characteristics has been a well-established research topic, particularly in the context of new wood production and quality control. Over the past decades, stationary optical measurement systems have been widely employed in industrial settings to monitor lumber surfaces for visible defects and structural irregularities. These systems aim to ensure consistent product quality by identifying surface anomalies early in the production line.

Extensive research has been dedicated to both general surface anomaly recognition (Louban, 2009; Syla et al., 2023) and the identification of specific features, such as wood knots (Ding et al., 2020; Yang et al., 2021; Chizhova et al., 2024), tree rings (Divya & Kaur, 2020), and surface cracks (Liu et al., 2020). However, these studies predominantly focus on newly processed timber under controlled imaging and lighting conditions.

In recent years, the application of machine learning and deep learning techniques has significantly advanced the field. For instance, He (2019) proposed a fully convolutional neural network (FCN) tailored to classify wood defects, effectively distinguishing between live knots, dead knots, and cracks using a customized dataset. Similarly, Fang et al. (2021) adopted the YOLOv5 architecture for real-time detection of surface knots in sawn timber, demonstrating the potential for integration into industrial automation pipelines.

1.3 Research Objectives and Contributions

Despite significant advances in automated wood surface analysis, there remains a critical gap in developing AI-based methods capable of reliably detecting wood knots and supporting structural assessment for historic timber structures under real-world conditions.

This study addresses this gap by constructing a dedicated data collection and developing a complete AI-based pipeline for the detection and assessment of wood knots in historic timber elements. The main contributions of this work are:

- The creation of large-scale, annotated datasets of historic timber surfaces and wood knots under diverse acquisition conditions.
- The integration of deep learning models into an end-to-end pipeline for automated knot detection and structural factor estimation, validated against expert measurements.

Through these contributions, we aim to demonstrate the potential of AI-assisted approaches in enhancing the structural evaluation and conservation of heritage timber structures.

2. Datasets Collection

2.1 Data Acquisition

Historic timber structures exhibit distinct visual and material characteristics such as surface weathering, discolouration, biological degradation and irregular geometry. These characteristics differ significantly from freshly processed or minimally aged timber, which is the focus of most existing datasets. Such datasets are inadequate for developing models tailored to the needs of heritage conservation, where accurate identification of wood knots and surface features is required under variable and often non-ideal conditions. The visual complexity of heritage timber requires specialised datasets that reflect natural wear, inconsistent textures and variations due to lighting. Existing datasets, such as the large-scale wood defect dataset by Kodytek et al. (2021), have been extensively used for automated vision-based quality control in timber production. However, our preliminary experiments on wood knot detection using yolov8m (Jocher et al., 2023) demonstrated that this dataset is not optimal for historical wood surfaces due to the differences between knots in fresh wood and those in aged timber.

To meet the specific needs of this study, we developed a custom image database of approximately 10,000 high-resolution images. These captures were systematically differentiated according to the following parameters

- Acquisition conditions: Images were captured under both controlled laboratory conditions and real-world environments, including low-light conditions, occlusions and limited accessibility typically found in historic buildings.
- Feature resolution: The database includes close-up images capturing fine structural features (e.g. knots, cracks) as well as wide-angle images of entire wooden beams to allow multi-scale analysis.
- Imaging technology: High-resolution RGB images through Nikon D6400, Nikon D850, Sony A7R IV; images through mobile devices (iPhone 13 Pro Max and iPhone 14 Pro) equipped with the 3DScanner app.
- Object positioning in architectural context: The database further distinguishes between accessible and embedded structural elements, considering variations in visibility and recording feasibility.

2.2 Preprocessing and Annotation

Following image acquisition, a multi-step preprocessing and annotation workflow was applied to ensure data quality and suitability for downstream AI training. As a first step, all collected images were reviewed for exposure-related artifacts: overexposed and underexposed images were automatically filtered using histogram-based thresholds. After this initial filtering, each image was manually inspected to ensure visual clarity, completeness of content, and the absence of capture errors such as motion blur or defocus. This ensured that only high-quality and representative samples were included in the whole data collection.

Dataset	Image amount	Resampled resolution	Task	Capture tool	Captured objects
Det-sf-v1	635	640 × 640	Detection	Nikon D850	dismantled timber elements and the old roof of a historic wooden house
Det-dominik-v1	606	640 × 640	Detection	Nikon D6400 & Sony V7R IV	roof structure of an outbuilding at the Dominican church in Bamberg
Det-dominik-v2	290	640 × 640	Detection	Nikon D6400	close-up images of knot features in the tower loft of the Dominican church in Bamberg
Det-dominik-v3	316	640 × 640	Detection	iPhone 14 pro	Roof structure in the tower loft of the Dominican church in Bamberg
Det-dominik-v4	246	640 × 640	Detection	Nikon D6400 & iPhone 14 pro	Main roof structure of the Dominican church
Det-mario-v1	307	640 × 640	Detection	iPhone 14 pro	Roof structure of the Marionette theatre
Seg-dominik-v1	584	1024 × 1024	Segmentation	iPhone 13 pro max	Roof structure in the tower loft of the Dominican church in Bamberg

Table 1. Summary of established fundamental datasets.

Figure 2. Samples from each fundamental dataset.

In the annotation phase, images were uploaded to the Roboflow platform and labelled according to task-specific requirements. For segmentation, polygon-based masks were created to distinguish *main beams* and *side beams*, allowing for structural separation in later processing. For detection, bounding boxes were applied to *wood knots* as the primary class. Although wood dowels, visually similar to knots, were also annotated during the process, all detection datasets used in this study were filtered to include only wood knots to ensure consistent evaluation of knot-specific model performance.

This task-specific annotation strategy ensures that the training data reflect real-world complexity while maintaining a controlled framework for evaluating segmentation and detection model performance. The final established fundamental datasets are summarized in Table 1 and several samples are illustrated in Figure 2.

3. Methodology

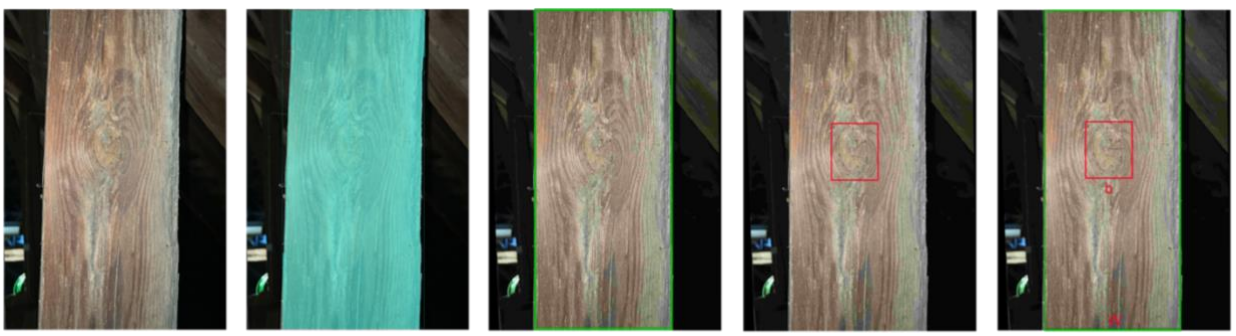
3.1 Workflow overview

The proposed workflow employs state-of-the-art deep learning architectures for timber surface segmentation and knot detection. The models are trained on the custom datasets under various experimental setups. The complete workflow comprises three main stages, as illustrated in Figure 3:

Stage 1 Timber Surface Segmentation:
The first stage involves instance segmentation to extract the target timber surface from the background structure. Several models like Detectron2 (Wu, 2019) and models from YOLO family are trained on the Seg-dominik-v1 dataset to perform this segmentation task. The selection criteria for the optimal model include segmentation accuracy, model size and inference speed, depending on the relevant practical acquisition in the future. The segmentation results are exported in both mask coordinate and PNG format.

Stage 2 Perspective adjustment:
Once the timber surface has been segmented, the largest instance by area is selected for perspective correction. This assumption is based on the observation that, under standard imaging conditions, the main surface occupies the largest contiguous area in the image. The extracted segmentation mask undergoes Douglas-Peucker (DP) simplification to reduce unnecessary complexity while preserving essential shape features. A perspective transformation matrix is then computed and applied to correct perspective distortions present in the original image. This step ensures that the output image is standardized and rectified, providing an optimal input for the subsequent wood knot detection. Correction of optical lens distortion is not taken into account in the current workflow, but will be considered in future implementations, particularly for on-device applications on mobile or edge platforms.

Stage 3 Wood Knot Detection:
In this stage, the perspective-corrected image is processed using a pre-trained detection model, which identifies all visible wood knots on the target timber surface. Currently, models from the YOLO family are employed due to their balanced detection efficiency and accuracy. In preliminary tests, several post-processing methods, such as k-Nearest Neighbours (kNN) feature validation, Non-Maximum Suppression (NMS), and Otsu’s thresholding, were evaluated to improve result consistency and bounding box refinement. Although these



The input image captured by iPhone

Timber surface segmentation through Stage 1

Perspective adjustment based on simplified segmented polygon through Stage 2

Wood knot detection through Stage 3

Factor calculation through shorter side of bounding box $\boldsymbol{b}$ and timber width $\boldsymbol{W}$ in pixel. For this example, Factor $\boldsymbol{A}$ = 0.3414

Figure 3. Visual illustration of an example processed through the proposed AI-assisted pipeline.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-267-2025 | © Author(s) 2025. CC BY 4.0 License.

methods have not yet been fully integrated into the final pipeline, they showed promising effects and are planned to be incorporated and systematically evaluated in future versions of the workflow.

Stage 4 Factor Estimation:
In the final stage, the adjusted timber surface polygon obtained from Stage 2 is combined with the bounding box of each detected knot from Stage 3. For each knot, the shortest side of its bounding box is measured and compared to the corresponding dimension (width or height) of the timber surface polygon. After evaluating all knots on a single timber element, the maximum of these individual ratios is taken as the final value of factor A, as defined above. This value characterizes the most critical knot influencing the grading of the timber element, as illustrated in Figure 1. Currently, only axis-aligned bounding boxes are supported, but future work will incorporate rotated bounding box adaptation for more accurate representation of angled knots.

With the adjusted timber surface mask and the refined bounding boxes of detected wood knots through the above three stages, the relative distance between each detected wood knot and the segmented timber boundary can be computed. This geometric relationship provides crucial insights into the structural integrity of the historic timber.

3.2 Model selections

To explore the architectural suitability of different deep learning models for timber surface segmentation and wood knot detection, we selected three representative model families: **YOLO**, **Detectron2** and **DeepLabV3** (Chen et al., 2018). These models were integrated into the workflow with distinct roles based on their design principles, task adaptability, and practical applicability.

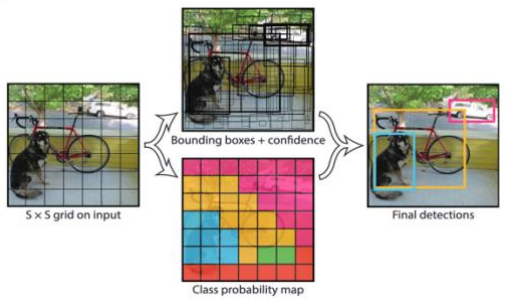


Figure 4. Grid-based object detection principle in YOLO, Redmon et al. (2016).

The YOLO (You Only Look Once) models, known for their one-stage architecture and high inference speed, were employed for both segmentation and detection tasks. The model series from YOLOv8 to YOLOv11 in YOLO family offers progressive improvements. YOLOv8 features a streamlined structure with decoupled heads and integrated segmentation branches, while YOLOv9 builds upon this by introducing Generalized ELAN (GELAN) and weight standardization, enhancing feature fusion and model convergence. YOLOv10 further advances the architecture with anchor-free detection, re-parameterization techniques, and conditional convolution layers. The latest iteration, YOLOv11, enhances backbone efficiency, reduces parameters by 22% compared with YOLOv8, and expands task support on Oriented Object Detection (OBB). Its optimized architecture balances accuracy and speed for edge deployment.

Despite these advancements, the core concept of YOLO detection remains consistent: the input image is divided into a grid, and each grid cell predicts bounding boxes and class probabilities. These predictions are then refined into final detections using confidence scores and non-maximum suppression. This process is visualized in Figure 4, which illustrates the flow from grid partitioning to final object localization.

In contrast, Detectron2 was employed exclusively for segmentation, offering a region-based instance segmentation pipeline built on the Mask R-CNN framework. Two model variants were selected: one using a ResNet-50 backbone with Feature Pyramid Networks (FPN) for a balance between accuracy and efficiency, and another with a deeper ResNet-101 backbone to support more detailed feature extraction and precise mask boundaries, albeit with greater computational cost. These models represent classical two-stage segmentation architectures, which are particularly effective in tasks requiring high spatial precision and the separation of multiple object instances.

As a semantic segmentation baseline, DeepLabV3 was included to evaluate the performance of encoder–decoder-based models. Three backbone configurations were tested: ResNet-50 for moderate accuracy with acceptable computational load, ResNet-101 for improved semantic depth at higher cost, and MobileNetV3-Large as a lightweight alternative more suited for edge deployment. These models were used to explore how well semantic segmentation performs in delineating timber surfaces under varying visual and environmental conditions.

4. Experiments

4.1 Experiment Settings

Based on the established datasets of historical wood structures for segmentation and detection tasks, and the overall pipeline designed for estimating wood knots on historical timber surfaces, the experiments aim to provide robust validation across various datasets and model architectures. The goal is to verify both the quality of the constructed datasets and the applicability of multiple state-of-the-art deep learning models.

Dataset	Data split train/val/test	Data Augmentation on train set	Abbrevi-ation
Det-sf-v1	445/95/95	1321	s1
Det-dominik-v1	426/90/90	1265	d1
Det-dominik-v2	204/43/43	607	d2
Det-dominik-v3	222/47/47	660	d3
Det-dominik-v4	174/36/36	517	d4
Det-mario-v1	215/46/46	640	m1
Seg-dominik-v1	408/88/88	1224	-

Table 2. Data split and augmentation for segmentation and detection experiments.

For clarity, we use short codes to refer to the datasets throughout the paper. Table 2 summarizes their full names, data splits, and the corresponding abbreviations used. For example, **d1** refers to *Det-dominik-v1*, and **m1** to *Det-mario-v1*. Combined datasets are referenced by merging abbreviations (e.g., **Det-d2d4m1** indicates a training set composed of Det-dominik-v2, v4, and Det-mario-v1). While dataset splitting was performed carefully, it remains possible that the same beam or knot appears in more than one subset if it was captured from different angles or under varying conditions during the data collection process.

In the Seg-dominik-v1 dataset for segmentation task, the training set was augmented using random hue and brightness shifts, with

additional Gaussian noise applied to up to 2% of the pixels in each sample.

For all training sets for detection task, augmentations included a combination of random changes in luminance, Gaussian blur, and additive noise, applied individually to each image. These procedures were implemented to increase the diversity of training samples and simulate realistic variation in visual conditions.

4.2 Segmentation models

According to the general pipeline introduced in Section 3.1, the main objective of the segmentation stage is to extract the primary structural regions of historical timber surfaces from their background and surrounding noise.

In the segmentation experiments, we selected models from three major architectures as mentioned: the YOLO family as representative single-stage models, Detectron2 for testing mask-based approaches, and the DeepLabV3 series for classical encoder–decoder segmentation. Based on our preliminary experiments results on segmentation models on dataset Seg-dominik-v1, all models in this stage were trained on the Seg-dominik-v1 dataset for 150 epochs under identical conditions and subsequently evaluated on the corresponding test set. The performance results on test set in terms of Mean IoU, Mean Dice, and inference time are summarized in Table 3.

Model	Mean IoU	Mean Dice	Inf. Time (ms)
yolov8m-seg	0.847	0.893	6.98
yolov9c-seg	0.859	0.906	10.49
yolo11n-seg	0.859	0.911	2.94
yolo11m-seg	0.859	0.909	7.25
yolo11l-seg	0.855	0.908	11.03
mask_rcnn_R_50_FPN_3x	0.789	0.842	131.29
mask_rcnn_R_101_FPN_3x	0.778	0.829	143.79
deeplabv3_resnet50	0.428	0.511	68.67
deeplabv3_resnet101	0.400	0.486	103.45
deeplabv3_mobilenet_v3_large	0.428	0.512	28.81

Table 3. Evaluation results of segmentation models on test set from Seg-dominik-v1.

The ***Mean IoU*** (Intersection of Union) and ***Mean Dice*** are used as the primary evaluation metrics to assess segmentation quality. The *mean IoU* can further be calculated based on the IoU between the *i*-th ground truth mask and its corresponding prediction in class *c*, where *C* represents the number of non-background classes (i.e., main beam and side beam in *Seg-dominik-v1*), N_c is the number of matched ground truth–prediction mask pairs in class *c*:

$$mean\ IoU = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{N_c}\sum_{i=1}^{N_c} IoU_{c,i}\right)$$

To fairly assess per-class performance, the *mean Dice* is computed as the average of Dice scores across all categories according to IoU. The Dice score reflects the spatial overlap between the predicted region and the ground truth and is particularly useful for evaluating segmentation performance on imbalanced datasets, as it gives more weight to correctly predicted regions than to the background.

Overall, YOLO-based segmentation models achieved the best performance on the Seg-dominik-v1 test set, with models like yolov9c-seg, yolo11n-seg, and yolo11m-seg combining high accuracy with inference times under 11 ms. Their superior performance likely stems from architectural suitability and better adaptation to small, domain-specific datasets. In contrast, DeepLabV3 models with heavier backbones (e.g., ResNet-101) typically require larger and more diverse data to generalize effectively, making them prone to underfitting or overfitting in this context. Given the identical training settings, the observed performance gap is primarily attributed to differences in model architecture rather than training configuration.

4.3 Detection models

4.3.1 Multi models training on fundamental datasets

In the first experiments step the chosen models mentioned in Section 3.2 from YOLO family for detection task were trained individually on the established fundamental datasets containing single class ***wood knots*** as shown in Table 2. All the models were trained with the identical hyperparameter setup with 150 epochs and Adam optimizer. To evaluate the performance of trained models there are mainly five metrics to evaluate:

- ***mAP50*** is the most commonly used performance metric, indicating the mean Average Precision (AP) at an IoU threshold of 0.5, measuring how well predicted boxes align with ground truth.
- ***mAP50-95*** extends mAP50 by averaging AP across IoU thresholds from 0.5 to 0.95 in steps of 0.05, which provides a more comprehensive assessment of localization quality.
- ***Precision*** is used to reflect the model's ability to avoid false positive samples, while ***Recall*** indicates how well it detects all true objects.
- The ***F1-score***, as the harmonic mean of mean precision and mean recall, provides a balanced metric that summarises overall recognition accuracy. It can be calculated as following:

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

The detailed values are likewise represented in the form of a heat map e.g. Figure 5 for F1-score. The results demonstrates that the datasets ***d1*** (Det-dominik-v1), ***d2*** (Det-dominik-v2), ***d4*** (Det-dominik-v4), and ***m1*** (Det-mario-v1) generally lead to higher F1-scores across most YOLO models, indicating that these datasets

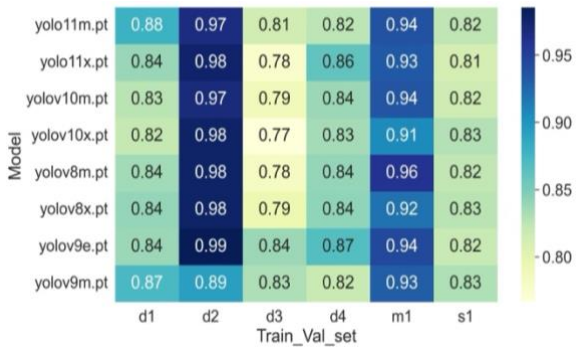


Figure 5. F1-score heatmap from from validation results: Each cell shows the F1-score of a particular model *Y* (rows) trained and validated on the dataset *X* (columns). These scores were recorded during training and reflect model performance on their respective validation sets.

offer more consistent training conditions or contain more representative features for model learning. Conversely, the performance on ***d3*** (Det-dominik-v3) and ***s1*** (Det-sf-v1) tends to be relatively lower or more variable, suggesting possible domain differences or increased dataset complexity.

4.3.2 Cross-validations on fundamental datasets

To assess the generalization capability of each fundamental dataset, we conducted cross-validation tests using multiple YOLO models trained on the same dataset shown in Figure 5. For each training set, the models were evaluated on the test sets of all other datasets. The average F1-scores across models were then calculated to obtain a robust estimation of how well a training dataset supports cross-domain detection. These results are visualized as a heatmap in Figure 6.

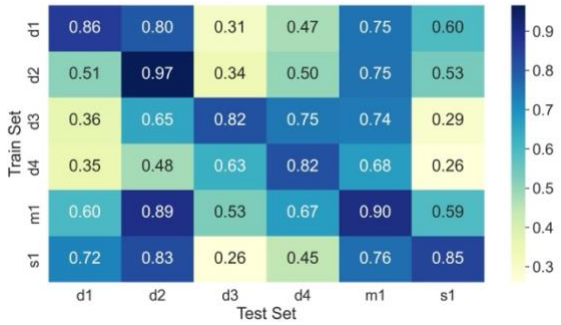


Figure 6. Average F1-score heatmap from test results: Each cell represents the average F1-score of all models trained on the training set of dataset *Y* (rows), evaluated on the held-out test set of dataset *X* (columns). The scores are averaged over all test images.

Based on the cross-dataset evaluation results shown in Figure 6 and the previous training results in Figure 5, the datasets ***d2***, ***d4***, ***m1*** and ***s1*** were selected as the base datasets for the next phase of mixed-dataset training and validation. This decision was motivated by three key factors:

- *In-domain performance*: All those datasets exhibited strong F1-scores when models were trained and tested on the same dataset, indicating clear and learnable feature patterns as well as reliable annotations.
- *Cross-domain generalization*: Compared to other datasets, models trained on *d2*, *m1* and *s1* exhibited relatively balanced and consistent F1-scores across multiple unseen domains (off-diagonal cells in Figure 6). Although *d4* demonstrated less consistent cross-domain generalization than *d3* (Figure 6), it was selected based on its highly stable in-domain performance during training (Figure 5), which reflects well-structured features and consistent annotations. These properties are essential for establishing robust model foundations in mixed-domain training.
- *Diversity and complementarity*: These datasets differ in content structure, defect patterns, and source domains (e.g., Dataset from Dominican church vs. from Marionette theatre shown in Figure 2). This diversity provides complementary information and broader feature coverage, which helps build models that generalize better across varied and challenging scenarios.

Although not selected as primary sources in the initial mixed-domain training phase, datasets such as d1 and d3 will be revisited in subsequent research stages. Their distinct domain characteristics and promising cross-domain signals (as observed in Figure 6) may offer additional insights into model adaptability under diverse deployment conditions.

4.3.3 Combinational Dataset Experiments

To assess the impact of data diversity on model generalization, we constructed several mixed training sets by combining the augmented training data from selected fundamental datasets (e.g., d2, d4, m1, s1), while retaining each dataset's original validation and test splits to ensure consistent evaluation. In addition to these combinations, three reference datasets were established: Det-dominik, containing only data from the Dominican church; Det-non-dom, from other sources; and Det-all, a fully mixed dataset aggregating all samples using the same augmentation and splitting strategy. This design enables training on diverse inputs while preserving domain-specific test conditions for cross-dataset comparison. A summary is provided in Table 4.

Mixed Dataset	Data split train/val/test	Shortcut	Resources
Det- d2d4	1124/79/79	d2d4	d2, d4
Det- d2m1	1247/89/89	d2m1	d2, m1
Det-d2m1s1	2568/184/184	d2m1s1	d2, m1, s1
Det- d4m1	1157/82/82	d4m1	d4, m1
Det- d2d4m1	1764/125/125	d2d4m1	d2, d4, m1
Det-non-dom	1961/141/141	mixAdd	m1, s1
Det-dominik	3049/216/216	mixDom	d1, d2, d3, d4
Det-all	5010/357/357	mixAll	d1, d2, d3, d4, m1, s1

Table 4. Overview of combinational datasets used in cross-domain training experiments. (Each dataset is constructed by combining the augmented training sets and original validation/test sets of multiple fundamental datasets for detection as listed in the Table 2.)

Three YOLO models，*yolov9e*, *yolo11m*, and *yolo11x*， were selected for this phase based on their consistently strong performance in previous evaluations, making them well-suited for further testing under mixed training conditions. All models were trained on the above mixed datasets for 150 epochs using the same hyperparameter setup as in the previous single-dataset experiments. Due to the architectural complexity of the *yolov9e* model, training on the *Det-all* dataset resulted in excessively long durations and persistent errors. As a result, this configuration was excluded from the final evaluation. Figure 7 presents the F1-scores of trained YOLO models evaluated on the validation set derived from their respective training data combinations, reflecting the model's fitting performance and early-stage generalisation within similar domains. As shown in the figure, the combinations involving datasets ***d2***, ***d4***, and ***m1*** (e.g., ***d2d4***, ***d2d4m1***, and ***d2m1***) consistently yield the highest F1 scores across all three models during the training stage, which confirms the generalisation potential of these data sources identified in earlier experiments.

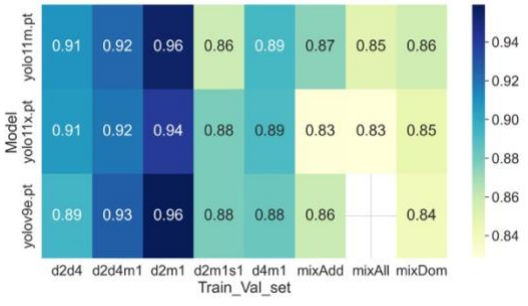


Figure 7. F1-score heatmap from trained models on various mixed datasets.

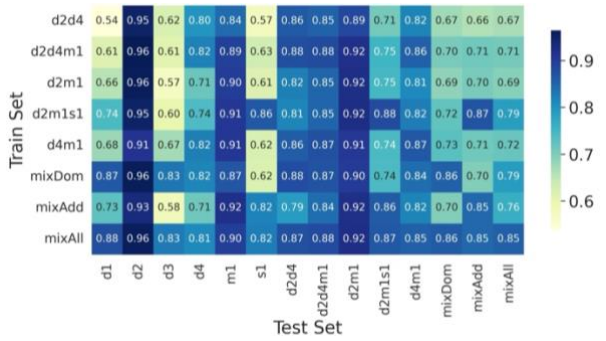


Figure 8. Average F1-score heatmap from test results on each test set of single datasets and mixed datasets across all three trained models in Figure 7.

The experimental results in Figure 8 reveal a clear trade-off between in-domain accuracy and cross-domain robustness. Specifically, when used individually or in limited combinations (e.g. *d2d4*, *d2d4m1*), datasets such as *d2*, *d4* and *m1* achieve exceptionally high F1-scores when evaluated in their respective domains. This suggests that models trained on these datasets can accurately capture domain-specific features. In contrast, mixed datasets such as *d2m1s1* demonstrate superior generalisation across diverse test sets, consistently yielding balanced and stable F1 scores. While their peak performance on any single domain may be slightly lower, their robustness to domain shifts makes them more suitable for real-world applications where unseen environments are common. Therefore, this comparison highlights the strategic choice between training for precision in targeted domains and achieving robust generalisation for broader applicability.

5. Evaluation

To enable an objective and independent evaluation of the system, we first selected two appropriate models, *yolo11m-seg* trained on *Seg-dominik-v1* for segmentation and *yolo11m* trained on *mixAll dataset* (all training samples) for detection, which have a notable performance based on the results of the previous experiments. In collaboration with conservation experts, we manually measured the largest visible knot factor on various wood elements as reference data. The evaluated timber structure had not been part of any previous training and validation datasets, which ensures an unbiased assessment.

Timber number	Manuel measured factor	Factor from AI pipeline	Relative error
Timber 1	0.0625	0.0956	52.96%
Timber 2	0.0652	0.1021	56.55%
…	…	…	…
Timber 37	0.4118	0.4716	14.51%
Timber 38	0.4130	0.3421	-17.18%
Timber 39	0.4375	0.2940	-32.80%
Timber 40	0.4583	0.5037	9.90%

Table 5. Sample-wise difference between AI-estimated greatest knot factors and manually measured knot factors.

In parallel, we randomly captured more than 900 images covering 40 timber elements using an iPhone 13 pro max, with an average of over 20 images per timber. These images cover various knot-bearing regions and different positions along each timber element. Then we processed them through the AI pipeline, which automatically segmented the timber surfaces, detected the wood knots, and calculated the knot factor for each image. For each timber element, we calculated the knot factor from all associated images and selected the maximum value as the final estimate, following the conservative principle of visual grading. The resulting values were used for comparison with expert measurements, as shown in Table 5.

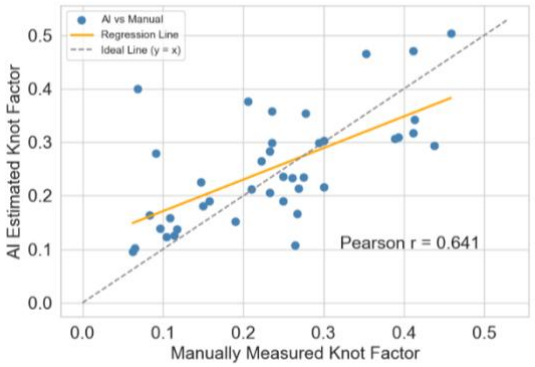


Figure 9. Correlation analysis of knot factors derived from manual measurement and AI-based estimation.

Across the 40 test timber elements, the AI pipeline successfully segmented most of the main surfaces that contain visible wood knots, and detected at least one knot per sample, allowing for the estimation of the corresponding maximum knot factor. However, a few surface segments could not be accurately extracted due to limitations in image quality and acquisition angles. These segmentation errors affected the subsequent knot detection and hindered the accuracy of factor estimation. Figure 9 illustrates the correlation between manually measured and AI-estimated knot factors across the 40 timber elements listed in Table 5. A Pearson correlation coefficient of 0.641 indicates a moderate positive relationship between the two measurements.

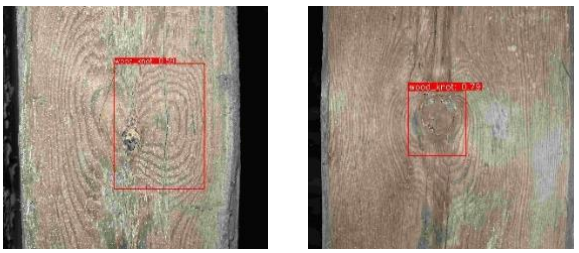


Figure 10. Samples illustrating false knot detection (left) and bounding box overestimation (right).

Compared to the manually measured knot factors, the AI-derived factors exhibited an average relative error of 33.13% (excluding outliers) and a median relative error of 25.82%, which indicates a generally good alignment in terms of geometric accuracy. Upon detailed review, two key types of error were identified, as shown in Figure 10:

- *False knot detection*: In two timber samples, the system mistakenly identified tree rings as knots, leading to an incorrect factor calculation. This type of misclassification suggests a need for improved differentiation between knots and ring-like structures, potentially through KNN filtering or similar post-processing techniques, which may help increase detection accuracy in future iterations.
- *Bounding box overestimation*: In five other timbers, the knot itself was correctly detected, but due to its small size, the bounding box was significantly oversized (Figure 10). This overestimation may be seen as a mild Type I error in knot area evaluation and could be mitigated Otsu thresholding optimisation or improved NMS, which are under testing but not yet integrated into the current pipeline.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-267-2025 | © Author(s) 2025. CC BY 4.0 License.

These results highlight that the geometric size of the bounding box, relative to the actual knot size, is a critical factor influencing the accuracy of the AI pipeline, especially for low contrast knots or irregular textures. Bounding box refinement strategies will therefore be essential to further improve the accuracy of automated structure grading.

6. Conclusion

The key contributions of this research can be summarized as follows:

- Development of a complete AI-assisted pipeline as a proof-of-concept for segmentation, perspective correction, and wood knot detection on historic timber surfaces, enabling automated and interpretable structural assessment.
- Construction and feasibility study of large-scale, diversified datasets tailored to historic timber, combining data from multiple historic sites and imaging devices under real-world conditions.
- Demonstration of practical applicability through expert-validated evaluation on real timber elements, comparing AI-derived grading factors with manual measurements to assess real-world performance.

Although the current datasets cover diverse acquisition conditions and sites, including the Dominican church, they still reflect only a limited portion of historic timber variability. To enhance representativeness and model generalizability, future work will focus on expanding the dataset to include timber elements from a wider range of buildings and regions. In parallel, domain adaptation techniques will be explored to improve model robustness in unseen environments.

This AI-assisted workflow is intended for deployment on low-cost mobile or edge devices and will be tested by conservators and structural engineers under real-world conditions. Beside the improvements from current AI pipeline, future research will focus on:

- Evaluating detection accuracy through detailed analysis of knot size ratios (e.g., knot-to-width) against manual ground truth.
- Extending defect detection to cracks, insect boreholes, and surface degradation and other damage on timber surface.
- Integrating 3D data and georeferencing methods, such as LiDAR, Structure-from-Motion (SfM), and sensor-based measurements, for internal condition analysis and spatially accurate defect mapping in large-scale timber structures.

The implementation code and dataset used in this study are available at *https://woodknot-cipa25.github.io/*.

References

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.

DIN, 2003: DIN 4074-1:2003-06 – Strength grading of softwood – Part 1: Graded timber. Deutsches Institut für Normung (DIN), Berlin.

Ding, F., Zhuang, Z., Liu, Y., Jiang, D., Yan, X., Wang, Z., 2020: Detecting Defects on Solid Wood Panels Based on an Improved SSD Algorithm. *Sensors (Basel)*, 20(18), 5315.

Divya, K., Kaur, S., 2021: Dendrochronology with Deep Learning. *In:* 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, pp. 333–336. https://doi.org/10.1109/ICIEM51511.2021.9445305

Fang, Y., Guo, X., Chen, K., Zhou, Z., Ye, Q., 2021: Accurate and automated detection of surface knots on sawn timbers using YOLO-V5 model. *BioResources*, 16(3), 5390–5406.

Görlacher, R., Eckert, H., 1999: Historische Holztragwerke: Untersuchen, Berechnen und Instandsetzen. Sonderforschungsbereich 315, Universität Karlsruhe (TH), Karlsruhe.

He, T., Liu, Y., Xu, C., Zhou, X., Hu, Z., Fan, J., 2019: A Fully Convolutional Neural Network for Wood Defect Location and Identification. *IEEE Access*, 7, 123453–123462.

Jocher, G., Chaurasia, A., Qiu, J., 2023: Ultralytics YOLOv8. *GitHub repository*, https://github.com/ultralytics/ultralytics

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016, IEEE, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

Kodytek, P., Bodzas, A., Bilik, P., 2021: A large-scale image dataset of wood surface defects for automated vision-based quality control processes. *F1000Research*, 10, 581.

Liu, Y., Hou, M., Li, A., Dong, Y., Xie, L., Ji, Y., 2020: Automatic detection of timber-cracks in wooden architectural heritage using YOLOv3 algorithm. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 1471–1476. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1471-2020

Louban, R., 2009: Image Processing of Edge and Surface Defects: Theoretical Basis of Adaptive Algorithms with Numerous Practical Applications. Springer. https://doi.org/10.1007/978-3-642-00683-8

Chizhova, M., Pan, J., Luhmann, T., Karami, A., Menna, F., Remondino, F., Hess, M., Eißing, T., 2024: Towards automatic defects analyses for 3D structural monitoring of historic timber. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W4-2024, 103–110. https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-103-2024

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019: Detectron2. *GitHub repository*, https://github.com/facebookresearch/detectron2

Yang, Y., Wang, H., Jiang, D., Hu, Z., 2021: Surface Detection of Solid Wood Defects Based on SSD Improved with ResNet. *Forests*, 12(10), 1419. https://doi.org/10.3390/f12101419

Zhang, X., Sun, H., Xu, G., Duan, Y., Jan, J., Joris, J., Shi, J., 2024: Understanding the Effect of Knots on Mechanical Properties of Chinese Fir under Bending Test by Using X-ray Computed Tomography and Digital Image Correlation. *Forests*, 15(1), 174.