Multi-stage mask-aware Depth Enhancement for RGB–IR–stereo Fusion on historic Timber Surfaces

Junquan Pan[1], Maria Chizhova[1], Mona Hess[1], Thomas Luhmann[2], Ferdinand Maiwald[3],

[1] Digital Technologies in Heritage Conservation, Centre for Heritage Conservation Studies and Technologies (KDWT), University of Bamberg, Bamberg, Germany – (junquan.pan, maria.chizhova, mona.hess)@uni-bamberg.de
[2] Institute for Applied Photogrammetry and Geoinformatics (IAPG), Jade University of Applied Sciences, Oldenburg, Germany – luhmann@jade-hs.de
[3] Chair of Optical 3D-Metrology, Dresden University of Technology, Dresden, Germany – ferdinand.maiwald@tu-dresden.de

Keywords: Object detection, Instance segmentation, Mask-aware depth enhancement, RGB-IR–Stereo fusion, Heritage timber, Damage assessment.

Abstract

This paper presents a mask-aware multi-stage depth enhancement framework for digital documentation of historical timber surfaces using RGB–Stereo-IR fusion. Accurate geometric recording of aged wood features such as wooden knots remains challenging due to uneven illumination and weak texture. The proposed pipeline, which aims to stabilise depth geometry under uneven illumination and low-texture surface conditions, integrates object detection, instance segmentation and confidence-guided depth refinement across three stages: (A) TV(total variation)-regularized mask-aware refinement, (B) confidence-weighted multi-view fusion, and (C) patch-based stereo reconstruction. Experiments on historical timber beams under varying illumination demonstrate improved depth completeness and geometric consistency, achieving a residual standard deviation below 0.6 mm in bright scenes and stable reconstruction in low-light conditions. The framework offers a practical solution for depth reconstruction of cultural heritage timber, supporting more reliable feature detection and analysis.

1. Introduction

1.1 Background and Motivation

Estimating the structural integrity of aged timber beams is essential for restoration planning and load-bearing assessment. Traditional assessment methods are often destructive; consequently, non-invasive digital and AI-assisted documentation techniques are becoming increasingly important in heritage conservation. According to DIN 4074-1 (Deutsches Institut für Normung, 2003), the distribution of knots on timber surfaces is an important indicator for strength grading of softwood. This property can potentially be analysed automatically using machine learning or AI-based methods.

Recent technologies in 3D recording, such as Structure-from-Motion (SfM), terrestrial LiDAR, and active stereo sensing, have improved heritage documentation precision. However, accurately digitising aged wood remains challenging due to uneven illumination, surface darkening, and anisotropic reflectance. Traditional RGB imaging often fails to capture subtle texture and geometry, especially under low-light or glossy conditions. This has led to the integration of active infrared (IR) stereo imaging technology to enhance the capture of geometric details and facilitate structural analysis

1.2 Challenges in the Documentation and Analysis of Historic Timber Surfaces

Despite its potential, combining RGB data with stereo infrared data still presents a number of challenges: (1) Timber surface exhibits distinct reflectance properties in visible and IR spectra. This can result in inconsistent appearances and textures across different modalities. (2) Repetitive grain patterns, low-contrast regions and specular reflections can lead to erroneous disparity estimation and depth discontinuities. (3) Small calibration inaccuracies between the RGB and IR sensors can result in noticeable spatial misalignment, particularly around damaged regions such as knots. (4) Cross-modal discrepancies can further undermine photometric and geometric consistency. Together, these factors hinder accurate surface reconstruction and the reliable identification of features such as knots, cracks, and tool marks.

1.3 Contribution

Rather than addressing the full structural assessment directly, this work focuses on improving the robustness and geometric consistency of close-range timber surface documentation. This provides a more reliable basis for subsequent feature analysis and condition interpretation. This study proposes a **mask-aware, confidence-guided depth refinement framework** for the multi-modal reconstruction of historic timber surfaces. The framework progressively enhances depth consistency and completeness by integrating structural and photometric cues from RGB and IR modalities through a three-stage refinement strategy. Its main contributions are as follows:

(1) **A unified multi-modal pipeline** combining RGB and stereo-IR imaging for robust surface reconstruction under varying illumination.
(2) **A confidence-weighted α-fusion mechanism** that adaptively integrates depth cues based on local reliability and cross-view consistency.
(3) **A three-stage enhancement process** including (A) TV-regularised mask-aware refinement, (B) confidence-based depth fusion, and (C) patch-level stereo reconstruction for fine geometric recovery.

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-XI-1-2026-217-2026 | © Author(s) 2026. CC BY 4.0 License.

(4) **A quantitative evaluation protocol** to assess mask quality, depth quality, and depth reliability across enhancement stages using real historic timber data, evaluated using a set of well-defined performance metrics.

2. Related Work

2.1 Historic Timber Surface Analysis

Traditional approaches to the analysis of historic timber surfaces have primarily focused on visible surface traces, such as cracks and knots, and on their structural interpretation in timber assessment and conservation practice (Goerlacher et al., 1999, Xu, 2002). Recent research has demonstrated the structural importance of visible timber features, such as knots and cracks. These features provide useful evidence for estimating strength and grading. However, a purely visual assessment is insufficient for historic timber. More comprehensive 3D documentation and AI-assisted analysis are therefore required (Chizhova et al., 2024).

With the rapid development of modern AI techniques, recent studies have increasingly explored the use of machine learning and deep learning for timber damage analysis. Convolutional neural networks combined with image-processing techniques have been applied to classify wooden defects and quantify their characteristics from digital imagery (Ehtisham et al., 2024). In the context of historic timber, end-to-end workflows have further been developed for the documentation and automatic detection of wood knots in support of structural assessment (Pan et al., 2025). In addition, multiple nondestructive data sources have been integrated to evaluate defect conditions in wooden columns of ancient buildings (Li et al., 2025).

Nevertheless, existing timber analysis approaches remain limited when relying on single-source evidence for surface analysis. Therefore, combining multi-source information could represent the next step towards achieving a more robust analysis of historic timber surfaces, particularly in challenging lighting conditions and in the presence of complex material surfaces.

2.2 Stereo Depth Reconstruction

Depth estimation based on stereo images is the most common and physically grounded method for reconstructing the structure of 3D scenes. Traditional methods rely on hand-crafted features and explicit geometric constraints to compute dense disparity maps between rectified stereo pairs. Among these, Semi-Global Matching (SGM) and its refined variant SGBM remain widely adopted for their balance between accuracy and computational efficiency, achieved by performing cost aggregation along multiple 1D paths (Hirschmuller, 2005). To further improve disparity quality, particularly in regions with weak texture or illumination noise, regularized optimisation has been incorporated. Early work on regularization theory (Lagendijk et al., 1988) provided the foundation, while more recent filters such as Weighted Least Squares (WLS) smoothing offer edge-preserving refinement that produces sharp and stable disparity boundaries (Farbman et al., 2008).

However, despite their interpretability and geometric soundness, such methods often perform poorly when surface texture is minimal or lighting is inconsistent, which are conditions that are frequently encountered on aged or dark wooden surfaces.

2.3 YOLO and Segment Anything

While depth estimation captures the geometric structure of the surface, the semantic understanding of wood regions, such as knots, cracks, or biological markings, requires object-level and mask-level recognition. These semantic evidences can provide valuable priors to guide depth refinement toward the target regions. To this end, two representative frameworks are reviewed here as the semantic front-end of the proposed pipeline: the YOLO family and the Segment Anything Model (SAM). Their complementary capabilities in detection and segmentation enable a more comprehensive interpretation of historical timber surfaces beyond purely geometric reconstruction.

In this regard, the YOLO (You Only Look Once) series (Redmon et al., 2016) has redefined object detection through its single-shot, end-to-end formulation, unifying localisation and classification in a single efficient framework. The latest generation, YOLOv11 (Khanam and Hussain, 2024), further refines its backbone and neck architectures for improved accuracy and real-time performance, making it well-suited for mobile and on-site inspection of wooden structures.

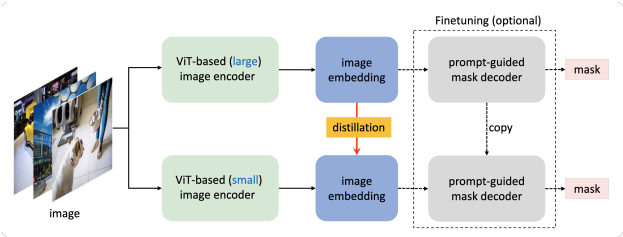


Figure 1. Decoupled distillation framework for MobileSAM (Zhang et al., 2023).

Complementary to detection, the Segment Anything Model (SAM) (Kirillov et al., 2023) introduced a foundational segmentation paradigm trained on over one billion masks, enabling general-purpose, prompt-based segmentation across domains. Its efficient variants, such as MobileSAM (Zhang et al., 2023) (Fig. 1), allow interactive or automated segmentation even on resource-constrained devices.

2.4 Multi-Modal Fusion

Multi-modal fusion uses complementary sensor data, such as LiDAR, RGB, infrared (IR) and thermal information, to improve understanding of scenes in challenging visual conditions. Integrating geometric, reflectance, and thermal information enhances the resilience of these methods in low-light, low-texture, or specular environments, such as those encountered during historical wood inspections.

In LiDAR-camera fusion, methods such as 3D-CVF (Yoo et al., 2020) and BEVFusion (Liang et al., 2022) have been developed to align cross-modal features into unified spatial representations (e.g. bird’s-eye view), thereby enhancing the accuracy and robustness of 3D detection. In the context of RGB-IR/thermal fusion, MEFNet (Lai et al., 2023) employs channel attention and soft weighting to adaptively emphasise modality-specific strengths during tasks such as segmentation.

In more recent work, researchers introduced a unified fusion framework (Xu et al., 2025) that identifies and reinforces reliable cross-modal depth cues under conditions of weak texture or ill-illuminated conditions. These findings emphasise the

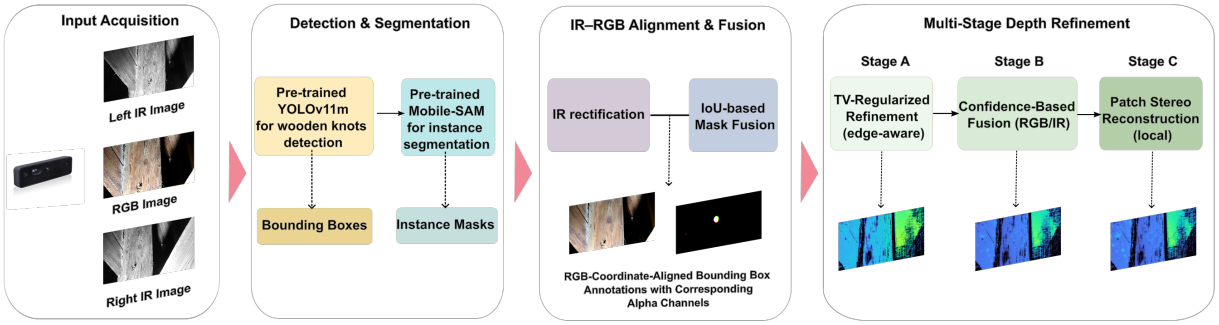


Figure 2. Overall workflow of the proposed AI-assisted multi-stage depth refinement framework. First, RGB and stereo-IR images are used for object detection and instance segmentation. This is followed by cross-modal IR–RGB alignment and fused mask generation. The depth map is then refined in three stages: (A) TV-regularised local refinement, (B) confidence-based multi-source fusion and (C) patch-level stereo reconstruction for low-confidence regions.

importance of integrating modality-specific features in order to stabilise depth estimation in visually degraded environments, which also offer valuable insights into the mask-aware RGB–IR refinement pipeline.

3. Methodology

To address the limitations of existing depth estimation methods under low-texture and low-light conditions, we propose an AI-assisted multi-stage refinement framework for depth enhancement on historical timber surfaces (Fig. 2). The proposed pipeline integrates RGB-IR acquisition using an active stereo camera, object detection, instance segmentation, geometric alignment, and mask-aware depth refinement into a unified workflow tailored to the characteristics of aged wooden materials. By combining active stereo sensing with mask-guided optimization, the method enhances both the completeness and stability of reconstructed depth maps, achieving consistent geometric recovery even under challenging illumination and material reflectance conditions.

3.1 OAK-D Pro Active Stereo Camera

The *OAK-D Pro* active stereo camera (*Luxonis*) was employed as the primary RGB-IR acquisition device. It integrates a pair of synchronised infrared sensors with a high-resolution RGB camera, which enables simultaneous capture of stereo IR and colour images. The system supports both active and passive stereo modes: In active mode, an IR dot projector tackles textureless surfaces for robust depth perception, while an IR flood illuminator ensures reliable operation in low-light or no-light conditions.

The device performs hardware-accelerated stereo matching and on-board depth computation within an optimal operating range of approximately 0.7-12 m. In addition to the original RGB and IR images, both the in-device disparity map and depth map can be exported for further processing. For close-range experiments on timber surfaces, the acquired image resolution was reduced to extend the effective capture range to around 0.4 m. An example scene captured without artificial illumination is shown in Figs. 3a, 3b, and 3c (RGB, left IR, and right IR).

3.2 Object Detection and Instance Segmentation

The proposed workflow employs RGB-IR image pairs captured by the OAK-D Pro stereo camera as input for semantic analysis of historical timber surfaces. Object detection is performed using a fine-tuned YOLOv11m model, which is optimised for identifying wooden knots under varying texture and illumination conditions. Following detection, instance segmentation is carried out using the Mobile-SAM framework, which balances computational efficiency and segmentation accuracy.

For each detected object, bounding boxes and segmentation masks are extracted from the RGB and stereo-IR images, ensuring consistent instance correspondence across modalities. These outputs serve as semantic priors for the subsequent alignment and depth refinement stages. The resulting segmentation mask is presented in Fig. 3d.

3.3 IR–RGB Alignment and Rectification

To achieve geometric consistency across modalities, we first perform stereo rectification of the left and right IR images and the corresponding masks using the calibrated intrinsic and extrinsic parameters obtained from exported camera calibration of OAK-D Pro. This process undistorts and rectifies both infrared images onto a shared epipolar plane (Luhmann et al., 2023), allowing pixel-wise correspondence to be established along the horizontal direction. The rectified masks are then reprojected into the RGB coordinate frame based on the extrinsic transformation ($T_{\mathrm{L}\to\mathrm{RGB}}$). This ensures depth-consistent alignment across all views. Figs. 3e and 3f show the mask aligned to the RGB frame and the corresponding bounding boxes on the RGB image.

Geometric consistency between the reprojected masks from different views is then quantified using the intersection-over-union ($\mathrm{IoU}_{L,R}$) metric as follows:

$$IoU_{(M_1,M_2)} = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2| + \epsilon} \tag{1}$$

where $\epsilon > 0$ is a small constant for numerical stability.

The corresponding intersection-over-union (IoU) values between masks from the left and right infrared images ($\mathrm{IoU}_{\mathrm{L,R}}$), as well as between the RGB image and each infrared view ($\mathrm{IoU}_{\mathrm{RGB,L}}$, $\mathrm{IoU}_{\mathrm{RGB,R}}$), are computed to evaluate the consistency of the re-projection results.

The resulting multi-view fusion mask $\tilde{M}$ combines RGB and IR masks with IoU-based attention factors to maintain geometric

alignment, which can be obtained through:

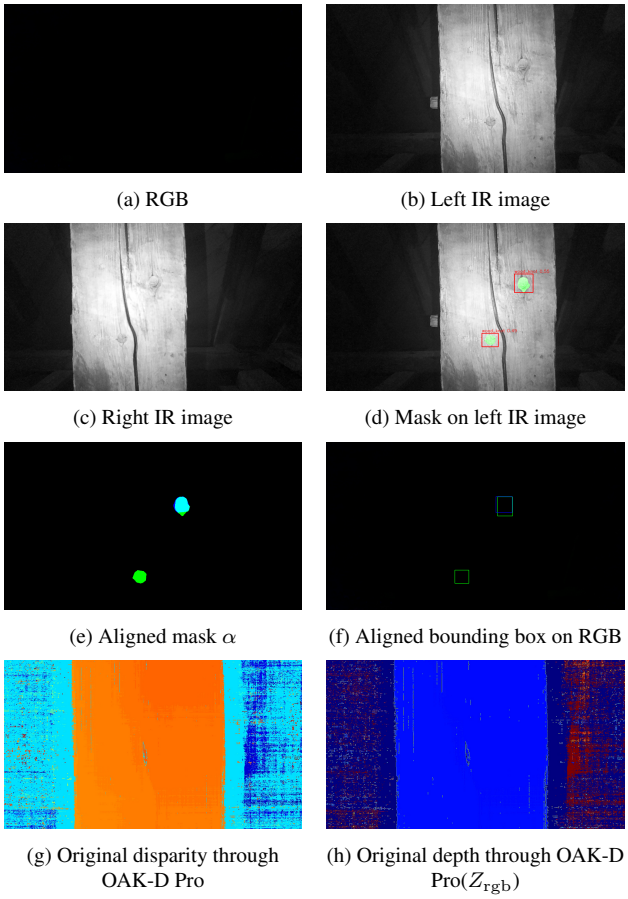
$$\tilde{M} = \gamma_{\text{rgb}}\, A_{\text{rgb}}\, M_{\text{rgb}} + \gamma_L\, A_L\, \alpha\, M_L + \gamma_R\, A_R\, (1-\alpha)\, M_R \quad (2)$$

where γ_{rgb}, γ_L, and γ_R are global weighting coefficients normalized such that $\gamma_{\text{rgb}} + \gamma_L + \gamma_R = 1$. In this research, the default weighting is $\gamma_{\text{rgb}} : \gamma_{\text{L}} : \gamma_{\text{R}} = 0.5 : 0.25 : 0.25$. The dominance factor α adaptively weights the left and right IR views based on their structural and coverage consistency. The IoU-based attention factors A_{rgb}, A_L, A_R capture geometric overlap between different modalities:

$$\begin{aligned} A_{\text{rgb}} &= 0.5\big(IoU(M_L, M_{\text{rgb}}) + IoU(M_R, M_{\text{rgb}})\big) \\ A_L &= 0.5\big(IoU(M_L, M_R) + IoU(M_L, M_{\text{rgb}})\big) \\ A_R &= 0.5\big(IoU(M_L, M_R) + IoU(M_R, M_{\text{rgb}})\big) \end{aligned} \quad (3)$$

3.4 Multi-Stage Depth Enhancement

After multi-view mask alignment, the depth enhancement proceeds through three stages: A-level refinement, B-level confidence fusion, and C-level patch reconstruction. The purpose of this stage-wise process is to progressively improve completeness and geometric consistency of the depth map. Representative results of the three-stage enhancement process are shown in Fig. 4.



(a) RGB (b) Left IR image (c) Right IR image (d) Mask on left IR image (e) Aligned mask α (f) Aligned bounding box on RGB (g) Original disparity through OAK-D Pro (h) Original depth through OAK-D Pro(Z_{rgb})

Figure 3. RGB and infrared (IR) images of historical timber surface (d_4) captured under low-light conditions, with corresponding detection, segmentation, and depth processing results.

A-level: TV-Regularized Mask-Aware Refinement. The first stage locally optimizes the raw depth Z_0 within fused mask regions $\tilde{M}$ using a total variation (TV) regularization framework (Rudin et al., 1992), which can be practically instantiated as a pixel-wise refinement loss that explicitly incorporates photometric consistency and edge-awareness as following:

$$\begin{aligned} \mathcal{L}_{\text{refine}} =& \mathcal{L}_{\text{p}} + \mathcal{L}_{\text{g}} + \mathcal{L}_{\text{e}} \\ =& \quad \lambda_p \sum_i m_i \; |I_L(i) - I_R(\pi(i, z_i))| \\ &+ \lambda_g \sum_i m_i \; \|\nabla z_i\|_1 \\ &+ \lambda_e \sum_i (1 - E_i)\, |\nabla z_i| \end{aligned} \quad (4)$$

where I_L and I_R are the rectified infrared images, $\pi(i, z_i)$ is the disparity-based correspondence, $m_i \in \tilde{M}$ is the fused mask indicator, and E_i denotes the edge confidence derived from the RGB gradient map. $\nabla z_i = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right)_i$ denotes the spatial gradient of the depth map at pixel i, which is used to regularize local smoothness while preserving structural discontinuities. The weights $(\lambda_p, \lambda_g, \lambda_e)$ balance the photometric, smoothness, and edge-preserving terms.

This balances data fidelity and smoothness while preserving edges inside $\tilde{M}$. In practice, the optimization also integrates photometric consistency between rectified IR pairs and an edge-aware weighting derived from the RGB gradients, yielding depth maps that remain smooth in homogeneous areas but respect visible boundaries.

B-level: Confidence-Based Depth Fusion. The second stage integrates multiple depth sources, namely the camera-based Z_{rgb}, the stereo-projected $Z_{\text{stereo}\to\text{rgb}}$, and the patch-based $Z_{\text{patch}\to\text{rgb}}$, through a confidence-weighted blending process:

$$Z_{\text{fused}} = \frac{w_0\, Z_{\text{rgb}} + w_1\, Z_{\text{stereo}\to\text{rgb}} + w_2\, Z_{\text{patch}\to\text{rgb}}}{w_0 + w_1 + w_2 + \epsilon}, \quad w_i \propto \text{conf}_i \quad (5)$$

Each confidence conf_i reflects both mask quality (Edge-F_1, coverage) and local image gradient strength, defined as:

$$\text{conf}_i = \lambda_1\, F_{1,i} + \lambda_2\, \text{cov}_i + \lambda_3\, G_i \quad (6)$$

where $F_{1,i}$ and cov_i denote the edge-based F_1 score and mask coverage for the i-th modality, respectively, and G_i represents the normalized gradient magnitude within the corresponding region of Z_i. The weighting coefficients λ_1, λ_2, and λ_3 control the relative influence of geometric and photometric cues.

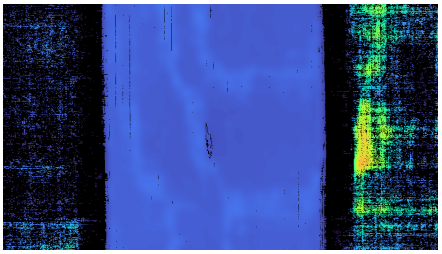
C-level: Patch-Based Stereo Reconstruction. Building upon the confidence-fusion framework of the B-level stage, the third stage focuses on regions with low fused confidence or missing depth values. Instead of global fusion, it performs **local stereo reconstruction** within and around uncertain mask regions to recover fine-scale geometry and fill unreliable depth zones, ensuring smooth transitions between masked and unmasked areas.

Each low-confidence area is locally rectified and re-estimated using patch-wise disparity refinement. The reconstructed depth $Z_{\text{patch}\to\text{rgb}}$ is then reprojected to the RGB coordinate frame.

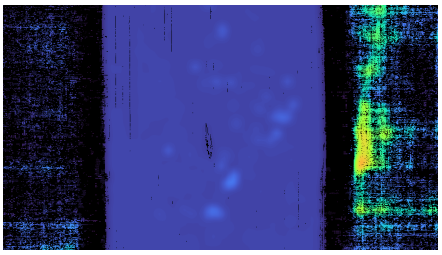
To ensure geometric stability, an adaptive update threshold $\delta_m \in [0.02, 0.05]\,\mathrm{m}$ (typically $0.05\,\mathrm{m}$) is applied:

$$|Z_{\mathrm{patch}\to\mathrm{rgb}} - Z_{\mathrm{rgb}}| > \delta_m \tag{7}$$

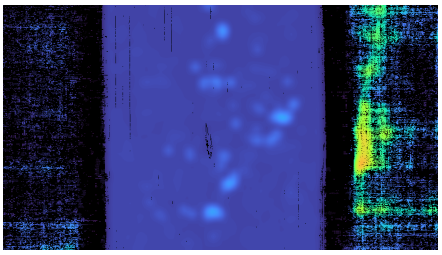
where only pixels exceeding this deviation are accepted as valid updates, while smaller deviations are suppressed as noise. The threshold δ_m is expressed in meters, since both Z_{rgb} and $Z_{\mathrm{patch}\to\mathrm{rgb}}$ denote metric depth values derived from the calibrated stereo baseline and focal length. The valid local reconstructions are subsequently reintegrated into the global map through the same confidence-weighted fusion rule in Eq. (5).



Stage A – TV-regularized refinement



Stage B – Confidence-based fusion



Stage C – Patch-based stereo reconstruction

Figure 4. Depth refinement results across the three enhancement stages (A–C) on historical timber surface from Fig.3.

3.5 Evaluation Metrics

To further quantitatively assess the quality of mask alignment and depth refinement, several indicators are computed to evaluate the completeness and edge consistency of the depth within each mask. There are three different evaluation categories which are defined as following.

3.5.1 Mask Quality Metrics. Edge-based F_1 score is computed to assess the geometric consistency and completeness of the masks projected onto the RGB reference plane.

Edge-F_1. The Edge-F_1 score (Eq. (8)) evaluates how well the predicted mask boundaries align with visible structural contours in the RGB image, which reflects the spatial accuracy of boundary alignment. A higher Edge-F_1 thus indicates a stronger alignment of the mask boundary with visible structural details in the RGB image.

$$\begin{aligned}
\text{Edge-}F_1 &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon} \\
\text{Precision} &= \frac{TP}{TP + FP + \epsilon} \\
\text{Recall} &= \frac{TP}{TP + FN + \epsilon}
\end{aligned} \tag{8}$$

In the above equations, TP, FP, and FN are determined through pixel-wise matching between the mask boundary $\partial\tilde{M}$ and the RGB edge map E_{rgb} obtained by the Canny operator:

$$\begin{aligned}
TP &= \left|\{\, p \in \partial\tilde{M} \mid \exists\, q \in E_{\mathrm{rgb}},\ \|p-q\|_2 < r \,\}\right| \\
FP &= \left|\{\, p \in \partial\tilde{M} \mid \nexists\, q \in E_{\mathrm{rgb}},\ \|p-q\|_2 < r \,\}\right| \\
FN &= \left|\{\, q \in E_{\mathrm{rgb}} \mid \nexists\, p \in \partial\tilde{M},\ \|p-q\|_2 < r \,\}\right|
\end{aligned} \tag{9}$$

where $\partial\tilde{M}$ denotes the set of edge pixels along the mask boundary, and E_{rgb} denotes the Canny-derived edge map of the RGB image. The radius r defines the maximum pixel distance for a match to be considered valid (typically $r =$ 2-3 in pixels).

Coverage. The coverage metric evaluates the completeness of the depth map within a given mask region. It measures the proportion of valid (finite and positive) depth pixels among all mask pixels, providing an indication of how well the region is covered by reliable depth estimates:

$$\begin{aligned}
\text{Coverage} &= \frac{N_{\text{valid}}}{N_{\text{mask}} + \epsilon} \\
N_{\text{valid}} &= \sum_i (m_i \wedge z_i > 0) \\
N_{\text{mask}} &= \sum_i m_i
\end{aligned} \tag{10}$$

3.5.2 Depth Quality Metrics. Depth residual variance, depth entropy and image-depth gradient correlation are employed to evaluate the smoothness, structural consistency, and completeness of the reconstructed depth within each mask.

Local depth residual standard deviation. The local depth residual standard deviation evaluates the geometric smoothness of the reconstructed surface within each mask. A local planar model $z = ax + by + c$ is fitted to the valid depth points, and the standard deviation of the residuals indicates the typical deviation of the measured depth from an ideal planar surface:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})^2}, \quad r_i = z_i - (ax_i + by_i + c) \tag{11}$$

where r_i denotes the residual error between the measured depth and the fitted plane, and $\bar{r}$ is the mean residual over all valid pixels.

Depth Entropy. The depth entropy quantifies the statistical dispersion of valid depth values within a masked region Ω. It is computed from the normalized histogram p_k of depth values

over K bins as:

$$H(Z;\Omega) = -\sum_{k=1}^{K} p_k \log(p_k + \epsilon) \tag{12}$$

where $p_k = h_k / \sum_i h_i$ and h_k is the frequency of bin k obtained from valid depth pixels within Ω. Higher entropy indicates greater surface irregularity or noise, whereas lower entropy corresponds to smoother and more stable depth reconstructions.

Image-Depth Gradient Correlation. To further assess whether depth edges follow RGB structure, we compute the Pearson correlation between gradient magnitudes:

$$\rho_\nabla(Z, I_{\mathrm{rgb}}; \Omega) \;=\; \mathrm{corr}(\|\nabla Z\|\,,\; \|\nabla I_{\mathrm{rgb}}\|)_{\mathbf{x}\in\Omega\cap\mathcal{V}} \tag{13}$$

Here $\|\nabla Z\| = \sqrt{(\partial_x Z)^2 + (\partial_y Z)^2}$ and $\|\nabla I_{\mathrm{rgb}}\|$ is computed on a grayscale version of I_{rgb}. Higher ρ_∇ implies better alignment of depth discontinuities with visible texture.

3.5.3 Stage-wise Change Metrics. To quantify the evolution of depth refinement across stages (A→B→C), three complementary measures are employed: median absolute depth change (ΔZ_{med}), depth structural similarity (SSIM_Z), and plane normal angle difference (θ_{plane}). All metrics are computed within the valid mask region $\Omega \cap \mathcal{V}$.

Median absolute depth change. The stability between two stages $Z^{(p)}$ and $Z^{(q)}$ is measured by the median absolute depth difference:

$$\Delta Z_{\mathrm{med}}\left(Z^{(p)}, Z^{(q)}\right) = \mathrm{median}_{\mathbf{x}\in\Omega\cap\mathcal{V}} \left| Z^{(q)}(\mathbf{x}) - Z^{(p)}(\mathbf{x}) \right| \tag{14}$$

We report $\Delta Z_{\mathrm{med}}(A{\to}B)$, $\Delta Z_{\mathrm{med}}(B{\to}C)$, and $\Delta Z_{\mathrm{med}}(A{\to}C)$, where smaller values indicate more stable and convergent depth estimates.

Depth SSIM. To assess structural similarity between two depth maps $Z^{(p)}$ and $Z^{(q)}$, we adapt the standard structural similarity index (SSIM) formulation (Wang et al., 2004) using a Gaussian window G_σ:

$$S_Z(\mathbf{x}) = \frac{(2\mu_p\mu_q + C_1)\,(2\sigma_{pq} + C_2)}{(\mu_p^2 + \mu_q^2 + C_1)\,(\sigma_p^2 + \sigma_q^2 + C_2)} \tag{15}$$

where μ_p, σ_p^2, and σ_{pq} are local means, variances, and covariances estimated via convolution with G_σ. The final score SSIM_Z is the spatial mean of $S_Z(\mathbf{x})$ over valid pixels, with higher values indicating stronger structural agreement.

Plane angle difference. Similar to local depth residual standard deviation, a local planar model of the form $z = ax + by + c$ is fitted to each depth map, and the orientation consistency between stages is measured as the angle between their normals:

$$\theta_{\mathrm{plane}}(Z^{(p)}, Z^{(q)}) = \arccos\left(\frac{\mathbf{n}^{(p)}\cdot\mathbf{n}^{(q)}}{\|\mathbf{n}^{(p)}\|\,\|\mathbf{n}^{(q)}\|}\right) \tag{16}$$

where $\mathbf{n}^{(p)} = (-a_p, -b_p, 1)$. A smaller θ_{plane} implies a higher geometric consistency across refinement stages.

4. Experiments

4.1 Object and Experimental Setup

A section of the timber roof structure of a centuries-old monastic church in southern Germany was selected for the purpose of evaluating the framework under realistic heritage conditions. The edifice is a protected cultural monument, and its roof assembly is constructed predominantly from softwood, a material frequently utilised in historical ecclesiastical architecture. Due to the age of the beams, they exhibit regions of both well-preserved and deteriorated surfaces, including darkened surfaces, weathering traces and pronounced knot structures. This provides a suitable setting for the testing of detection, segmentation and depth refinement algorithms on complex wooden heritage surfaces.

In total, eleven timber beams were documented under two illumination scenarios, namely with and without artificial lighting, to evaluate the robustness of the workflow under varying lighting conditions. A multitude of viewpoints incorporating wooden knots of various sizes and textures were carefully captured, resulting in the creation of 165 paired RGB–IR image sets. These image sets were then utilized for the purposes of depth refinement and cross-modal evaluation.

Acquisition Equipment. All image data were captured using the OAK-D Pro stereo camera (Luxonis) operated via the *DepthAI Viewer*. The RGB module recorded at a resolution of 1920×1080, while the stereo infrared (IR) pair captured synchronized images at 1280×720. On-board depth computation utilized **Left–Right Check** and **Extended Disparity** to ensure reliable stereo correspondence, with subpixel interpolation disabled to preserve sharp structural transitions and prevent smoothing of fine wood details.

Processing Pipeline. The captured RGB–IR pairs were processed by our multi-stage pipeline (Fig. 2). YOLOv11m and Mobile-SAM produced instance maps in both modalities, which served as semantic priors for geometric alignment between rectified IR stereo views and the RGB frame using calibrated parameters. We then applied the three-level depth refinement (Section 3.4) sequentially to each image–mask pair.

4.2 Evaluation Metrics

All evaluation metrics are aggregated across eleven timber beams. Since the number of captured samples varies among beams, each metric value represents the mean of all valid samples within the corresponding beam, with the error bars denoting the standard deviation across samples.

4.2.1 Mask Quality Evaluation The *coverage* metric (Fig. 5) evaluates the completeness of valid depth values within each mask region. Since nearly all pixels inside the mask regions contain valid depth values after preprocessing, the coverage remains close to 1.0 for all cases and is therefore not discussed further.

Under bright illumination (l*), the Edge-F_1 remains consistently high across all pairs (Fig. 6). The RGB channel achieves average F_1 scores of 0.7–0.8, indicating stable boundary alignment between IR and RGB masks and reliable module performance under sufficient contrast and texture.

In dark illumination (d*), nearly all Edge-F_1 scores drop to zero due to failed YOLO detections on underexposed frames, where

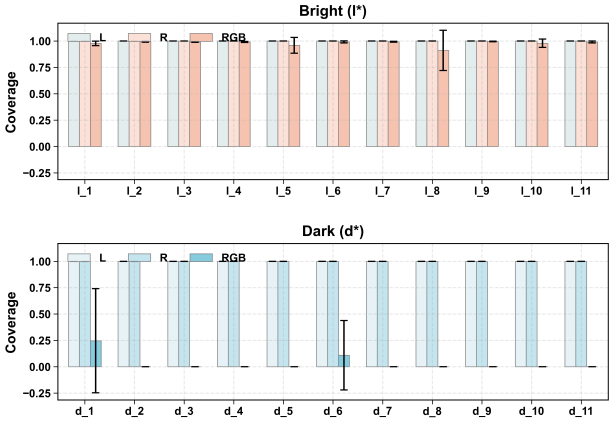


Figure 5. Mask coverage rate of A-level refinement across lighting conditions. The upper subfigure shows results under bright illumination, while the lower represents dark illumination.

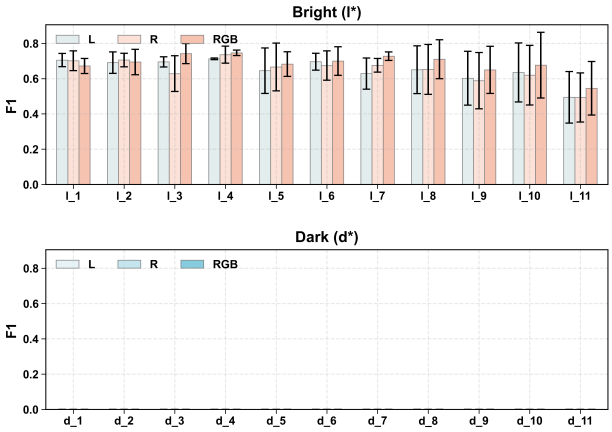


Figure 6. Edge-based F_1 score comparison under bright and dark illumination.

low contrast prevents valid object localisation. Consequently, Mobile-SAM cannot produce meaningful masks, revealing the strong dependence of mask-aware refinement on illumination quality.

4.2.2 Depth Quality Assessment The standard deviation of residuals (Fig. 7) indicates that the C-level reconstruction achieves sub-millimetre dispersion (< 0.6 mm) under bright illumination, suggesting that the preceding fusion and refinement stages effectively suppress local noise. Under dark conditions, the stereo-IR cues appear to remain the more reliable source of geometric information, and their residuals increase to around 0.6∼0.7 mm due to reduced IR texture contrast and noise from the active projection.

The Depth Entropy results (Fig. 8) further validate this trend: entropy is low and stable under bright light but fluctuates strongly in dark scenes. Low entropy signals confident depth surfaces with minimal noise, whereas high values signify uncertainty and error amplification.

Fig. 9 presents the gradient correlation between the reconstructed depth and RGB intensity maps across different illumination conditions and refinement stages. A clear divergence can be observed between bright and dark lighting scenarios.

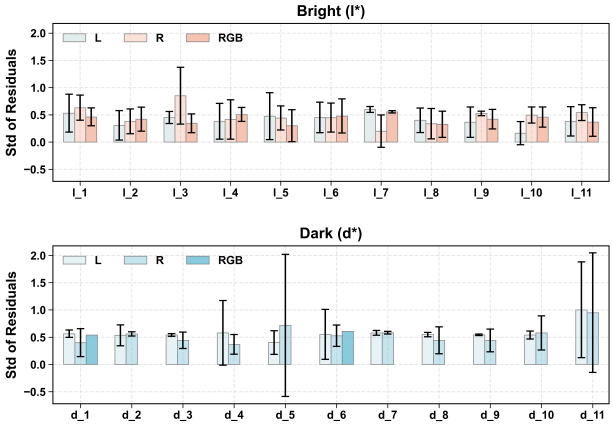


Figure 7. Standard deviation of residual depth errors (Std of Residuals) at the C-level reconstruction under bright and dark illumination. This is within the valid mask region at Stage C. It is reported in millimetres (mm). Lower values indicate smoother local geometry and less depth noise.

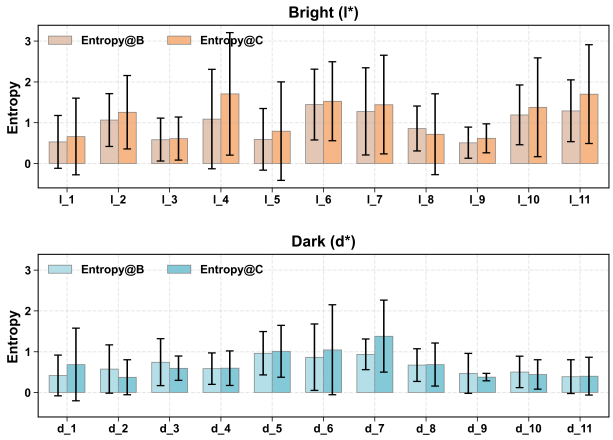


Figure 8. Entropy of depth distribution (Entropy@B & Entropy@C) under bright and dark illumination. Lower values indicate more concentrated and stable depth estimates, whereas higher values suggest increased irregularity or noise.

Under bright illumination (l*), the correlation remains generally low throughout all stages and even becomes slightly negative for several samples. This indicates that the intensity gradients in the RGB images, dominated by strong specular reflections and shading variations, do not correspond to actual geometric discontinuities. Although the A–C refinements enhance depth quality, they fail to improve gradient alignment because RGB gradients are driven by illumination effects instead of geometric structure.

In contrast, under insufficient illumination (d*), the active IR projection dominates surface illumination, producing consistent texture and edge cues that correlate well with true depth variations. The gradient correlation is generally higher (typically 0.4–0.7), although a slight decrease is observed from stage B to C, suggesting that the local patch reconstruction may introduce minor surface noise. This effect is partly attributed to the absence of RGB information during refinement, as the process relies solely on IR-based cues under low-light conditions.

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-XI-1-2026-217-2026 | © Author(s) 2026. CC BY 4.0 License.

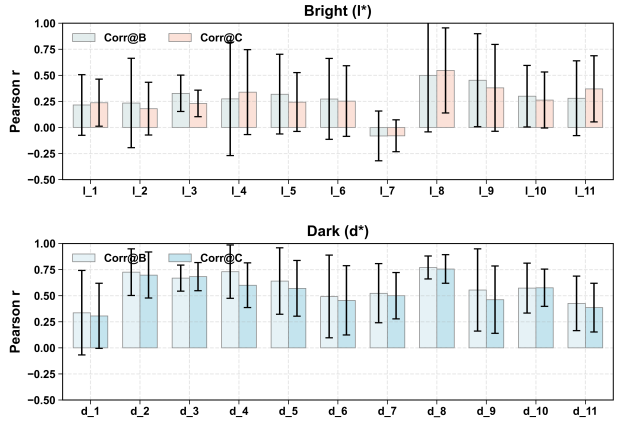


Figure 9. Gradient correlation (GradCorr@B & GradCorr@C) between RGB image gradient magnitudes and reconstructed depth gradient magnitudes under bright and dark illumination. Higher positive values indicate stronger alignment between the gradients of the RGB image and the reconstructed depth. Under dark illumination, the RGB gradient map becomes sparse and is dominated by a limited number of strong boundaries. This may artificially increase the correlation with reconstructed depth gradients.

4.2.3 Stage-wise Refinement Analysis The three-stage refinement pipeline (A → B → C) exhibits a consistent improvement trend across all pairs. In the following stage-wise analysis, the results are presented as aggregated metrics under bright and dark illumination conditions, without distinction between individual samples or beams.

Illumination	$\lvert\Delta Z\rvert$	A→B	B→C	A→C
Bright (l*)	Median	0.2932	0.0000	0.2952
	Min	0.0000	0.0000	0.0003
	Max	0.6321	0.0009	0.6332
Dark (d*)	Median	0.2903	0.0000	0.2934
	Min	0.2402	0.0000	0.2406
	Max	0.4322	0.0005	0.4332

Table 1. Stage-wise median depth deviation $\lvert\Delta Z\rvert$ (mm) under bright (l*) and dark (d*).

As shown in Table 1, the global fusion (A→B) performs most structural updates ($\lvert\Delta Z\rvert_{\text{median}} \approx 0.29$ mm), whereas the local patch refinement (B→C) yields minimal change due to thresholding and localized application, although local angular deviations can still become noticeable in a few dark cases. Under conditions of low illumination, the overall deviation magnitudes remain similar. This finding suggests that the global A→B correction remains the predominant factor, while the C-stage updates are subject to further limitations due to the reliance on IR cues, which are less effective in low-light conditions.

The SSIM values in Table 2 reveal a distinct contrast between stages. The A↔B transition exhibits low similarity (median 0.12), since Stage B substantially reorganises the depth structure during global fusion. In contrast, B↔C achieves high SSIM values (median 0.95 under bright and 0.80 under dark illumination), confirming that the C-stage mainly preserves the geometry produced in the previous stage and modifies only a limited set of pixels. Notably, some extreme SSIM values (e.g., >1000) occur for nearly uniform or invalid depth maps, where the variance terms in the SSIM formula become numerically unstable due to near-constant intensity. Overall, the results indicate that the multi-stage refinement primarily alters the global structure in the early stage and stabilises rapidly thereafter.

Illumination	SSIM	A↔B	B↔C	A↔C
Bright (l*)	Median	0.1173	0.9539	0.1108
	Min	-0.0202	-11.1219	-0.0235
	Max	0.6712	18.2520†	0.6468
Dark (d*)	Median	0.0302	0.7996	0.0318
	Min	-0.0219	-2.8347	-0.0361
	Max	0.2928	1706.1671†	0.2858

Table 2. Stage-wise SSIM (dimensionless) under bright (l*) and dark (d*).

The angular deviation between fitted planar regions (Table 3) remains low under bright illumination–typically below 0.3°–showing strong geometric consistency across stages. Under dark illumination, deviations increase modestly (median $\leq 0.8^\circ$, occasionally $> 1^\circ$) due to sparse or low-confidence depth, where normal estimation is noise-sensitive. In conclusion, the consistently low median deviations confirm the stability of the surface orientation and the robustness of the multi-stage refinement pipeline when subjected to varying lighting conditions.

Illumination	Angle(°)	A→B	B→C	A→C
Bright (l*)	Median	0.0269	0.0117	0.0325
	Min	0.0117	0.0115	0.0117
	Max	0.1843	0.0809	0.1843
Dark (d*)	Median	0.0369	0.0294	0.0759
	Min	0.0115	0.0115	0.0126
	Max	0.2413	1.0849‡	1.2521‡

Table 3. Stage-wise plane angle deviation (°) under bright (l*) and dark (d*).

4.3 Discussion

The experimental results confirm that the proposed workflow effectively integrates detection, segmentation, geometric transformation, and multi-stage refinement into a coherent reconstruction pipeline. Stage-wise metrics (Tables 1–3) show that the largest structural updates occur in the A→B fusion stage, while the C-level patch refinement stabilises the surface geometry with minimal angular deviations below 0.3°. The workflow thus validates a clear decoupling between global correction and local refinement.

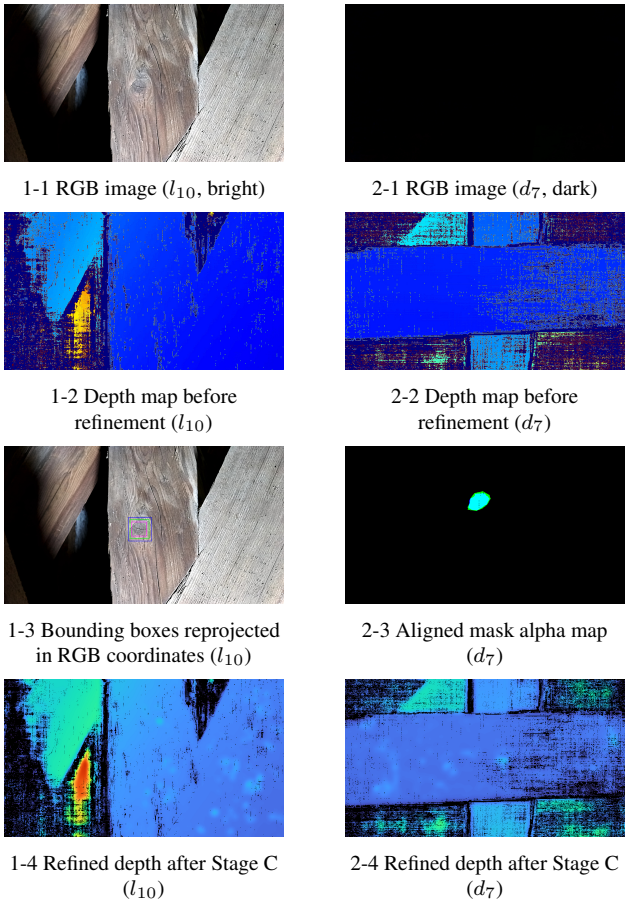
The results reveal a strong interdependence between illumination, mask reliability, and geometric alignment. Under sufficient lighting, accurate YOLO-SAM masks ensure stable RGB–IR alignment, allowing refinement modules to suppress noise and preserve geometry. In contrast, weak or uneven illumination reduces the reliability of RGB-based cues. Meanwhile, stereo correspondence in the IR images can still be affected by sparse texture, projection noise, or low-confidence regions, thereby propagating uncertainty into depth fusion.

While the visual difference between original and refined depth maps appears subtle (Fig. 10), the refinement targets local inconsistency reduction rather than depth amplification. Quantitatively, the lower residual standard deviation and entropy (Fig. 7, Fig. 8) indicate a more stable depth field without altering global structure.

It should be also noted that the test scenarios represent extreme lighting conditions, whereas real-world lighting is usu-

ally mixed or varies spatially, which can affect the reliability of RGB and the accuracy of the mask. Therefore, adapting mask thresholds and modality weights to local illumination would enhance robustness, particularly when transitioning between RGB- and IR-dominant regions.

Although the current pipeline prioritises robustness and geometric consistency, future work will also include runtime profiling and module simplification for practical on-site deployment.



1-1 RGB image (l_{10}, bright) | 2-1 RGB image (d_7, dark)
1-2 Depth map before refinement (l_{10}) | 2-2 Depth map before refinement (d_7)
1-3 Bounding boxes reprojected in RGB coordinates (l_{10}) | 2-3 Aligned mask alpha map (d_7)
1-4 Refined depth after Stage C (l_{10}) | 2-4 Refined depth after Stage C (d_7)

Figure 10. Further comparison of detection and refinement results of selected RGB-IR pairs for timber surfaces under bright (l_{10}) and dark (d_7) illumination.

5. Conclusion

This study presents a multi-stage, mask-aware depth enhancement framework that integrates RGB and stereo-IR imagery to more reliably document historical timber surfaces in 3D. Based on previous research into timber knot analysis and heritage data acquisition, this method addresses limitations commonly encountered in conventional close-range recording. It uses instance-aware masks, illumination-sensitive confidence cues, and staged refinement to improve depth completeness and geometric stability. The framework achieves sub-millimetre local residuals in favourable lighting conditions and maintains millimetre-level consistency in more challenging scenarios. Furthermore, the results demonstrate that alternative sensing modalities, especially active stereo IR, provide essential structural information for low-textured or degraded wooden surfaces, where RGB-only reconstruction is ineffective.

More broadly, this study demonstrates that the systematic combination of detection, segmentation, geometric transformation and depth refinement can improve the reliability of non-destructive surface documentation in heritage contexts. The study also highlights the importance of adapting to real-world illumination variability, suggesting that dynamic, sensor-aware thresholding is essential for practical implementation.

Future research will focus on developing **illumination-agnostic depth priors**, **cost-volume-based** RGB-IR correspondence learning and **cross-sensor fusion strategies** that can be applied to diverse heritage environments. These developments will pave the way for portable, multi-sensor inspection devices that can deliver consistent, on-site geometric documentation of wooden cultural heritage surfaces.

Acknowledgements

This work was supported by the Deutsche Bundesstiftung Umwelt (DBU) under the project *WoodF(ea)uture: Development of automated system for stability assessment of historic built-in timbers* (funding code: 39292/01).

References

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. XLVIII-2-W4-2024, 103–110.

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

Ehtisham, R., Qayyum, W., Camp, C. V., Plevris, V., Mir, J., Khan, Q.-u. Z., Ahmad, A., 2024. Computing the Characteristics of Defects in Wooden Structures Using Image Processing and CNN. 158, 105211.

Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R., 2008. Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation. *ACM Transactions on Graphics*, 27(3), 1–10.

Goerlacher, R., Falk, V. C., Eckert, H., 1999. *Historische holztragwerke. Untersuchen, berechnen und instandsetzen*.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2, 807–814 vol. 2.

Khanam, R., Hussain, M., 2024. YOLOv11: An Overview of the Key Architectural Enhancements.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment Anything.

Lagendijk, R., Biemond, J., Boekee, D., 1988. Regularized Iterative Image Restoration with Ringing Reduction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(12), 1874–1888.

Lai, W., Zeng, F., Hu, X., Li, W., He, S., Liu, Z., Jiang, Y., 2023. MEFNET: Multi-expert Fusion Network for RGB-Thermal Semantic Segmentation. *Engineering Applications of Artificial Intelligence*, 125, 106638.

Li, Y., Ouyang, W., Xin, Z., Zhang, H., Sun, S., Zhang, D., Zhang, W., 2025. Machine Learning for Defect Condition Rating of Wall Wooden Columns in Ancient Buildings. 22, e04458.

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z., 2022. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2023. *Close-Range Photogrammetry and 3D Imaging*. 4 edn, Walter de Gruyter, Berlin, 820 p.

Pan, J., Chizhova, M., Ebener, F., Luhmann, T., Ledig, C., Maiwald, F., Eißing, T., 2025. An End-to-End AI Pipeline for Wood Knot Detection to Enhance Structural Assessment in Historic Timber Structures. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-M-2-2025, 267–274.

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

Rudin, L. I., Osher, S., Fatemi, E., 1992. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4), 259–268.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Xu, J., Li, R., Cheng, K., Jiang, J., Liu, X., 2025. Unveiling the Depths: A Multi-Modal Fusion Framework for Challenging Scenarios. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, Atlanta, GA, USA, 6283–6290.

Xu, P., 2002. Estimating the Influence of Knots on the Local Longitudinal Stiffness in Radiata Pine Structural Timber. *Wood Science and Technology*, 36(6), 501–509.

Yoo, J. H., Kim, Y., Kim, J., Choi, J. W., 2020. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-view Spatial Feature Fusion for 3D Object Detection. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, 12372, Springer International Publishing, Cham, 720–736.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., Hong, C. S., 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications.