Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 3, September, 2024

A UAV Based Concrete Crack Detection and Segmentation Using 2-Stage Convolutional Network with Transfer Learning

Joses Sorilla [1], Timothy Scott C. Chu [1*], Alvin Y. Chua [1]

[1] *Department of Mechanical Engineering, De La Salle University, 2401 Taft Ave. Malate, Manila, Philippines.*

Received 29 April 2024; Revised 10 August 2024; Accepted 15 August 2024; Published 01 September 2024

Abstract

This study explores a non-destructive testing (NDT) method for crack detection using a two-stage convolutional neural network (CNN) model, incorporating a combination of AlexNet and YOLO models through transfer learning. Crack detection is pivotal for assessing structural integrity and ensuring timely maintenance interventions. The developed model was rigorously tested in simulated environments and through physical experimentations with the use of a UAV to evaluate its effectiveness. A 2-stage model, based on AlexNet and YOLO, was developed for crack classification and segmentation. The developed model leveraged transfer learning to address limitations from traditional CNN models. A known dataset was used to evaluate the developed model, benchmarking it against other models. The classification network achieved an accuracy rate exceeding 90%, while the segmentation network successfully identified and delineated cracks in 85.71% of the images. Finally, the developed model was deployed using a UAV to perform crack detection and segmentation in a controlled environment. These results underscore the model's proficiency in both detecting and segmenting structural cracks, highlighting its potential as a reliable tool for enhancing the maintenance and safety of architectural structures.

Keywords: Computer Vision; 2-Stage CNN; Crack Detection; Crack Segmentation; Transfer Learning; Unmanned Aerial Vehicles (UAV).

1. Introduction

Non-destructive testing (NDT) concrete crack detection is critical for maintaining the structural integrity, safety, and longevity of concrete structures. Cracks can compromise the load-bearing capacity of structures, leading to increasing the risk and maintenance costs if not addressed promptly. Traditionally, crack detection relies on manual visual inspections conducted by trained personnel. While this method is straightforward, it is time-consuming, subjective, and prone to human error. To overcome these limitations, advanced NDT methods such as ultrasonic testing, infrared thermography, and ground-penetrating radar (GPR), have been introduced. In recent years, automated methods using digital image processing have gained prominence, employing high-resolution cameras and algorithms to analyze concrete surfaces for cracks. Among these, machine learning-based techniques, particularly Convolutional Neural Networks (CNNs), have shown great potential in improving detection accuracy and efficiency [1]. Despite advancements in crack detection, existing methods often rely on complex, expensive image acquisition systems, limiting their feasibility for large-scale deployment [2, 3]. Many of these systems also face challenges in balancing real-time detection with precision, particularly in GNSS-denied environments [4, 5]. While CNN-based methods have improved detection accuracy, their dependency on numerous high-quality datasets and sensitivity to complex environments reinforce the need for more adaptable, efficient solutions.

* Corresponding author: timothy.chu@dlsu.edu.ph

http://dx.doi.org/10.28991/HIJ-2024-05-03-010

1.1. Review of Related Literature

For crack detection, researchers have explored both one-stage and two-stage models, often featuring an initial classification phase followed by segmentation [6]. One-stage models like YOLOv4 and YOLOv8 combine detection and localization in a single step, making them efficient for real-time monitoring of large infrastructure. For example, Li et al. (2024) [7] demonstrated that a ResNet50-based two-stage model with multilayer parallel residual attention (MPR) achieved a mean Pixel Accuracy (mPA) of 92.7%, a mean Intersection over Union (IoU) of 88.3%, and a processing speed of 36.5 FPS, suitable for real-time applications. Similarly, Paramanandham et al. (2023) [8] utilized Pixel Intensity Resemblance Measurement (PIRM) to enhance accuracy and robustness in crack analysis for structural applications, while a 2023 study showed that ResNet50 outperformed VGG16, VGG19, and MobileNet with a test accuracy of 99.88%, highlighting its strong generalization and fast convergence, particularly on smaller datasets [9, 10]. In contrast, two-stage models, such as the hybrid approach in this study, improve precision by focusing segmentation on relevant areas post-classification, which is particularly advantageous in complex or noisy environments [11, 12]. This can be seen in Yang et al. (2024) [13], where researchers developed an enhanced Mask R-CNN model for micro-crack detection on metal surfaces, refining feature extraction to detect small, intricate cracks even in low-contrast settings. Another study [14] introduced a two-stage framework using CNN and CTv2 networks for pixel-level pavement crack detection, achieving high accuracy across multiple datasets, with mean F1-scores up to 94.69%. While these models enhance detection accuracy, they are often constrained by computational demands and data requirements, a limitation addressed by transfer learning, which leverages pre-trained features to improve model performance [15].

Building on these insights, this study introduces a novel hybrid model that combines AlexNet for initial crack classification and YOLOv4 for segmentation, utilizing transfer learning to optimize accuracy with minimal data. This approach enhances both classification and segmentation efficiency, making it particularly suited for UAV-based applications where rapid, precise detection is essential. Notably, this specific combination of models has not yet been widely explored at the time of writing.

This paper is structured as follows: Section II, 'Theoretical Considerations,' outlines the foundational concepts pertinent to the study. Section III, 'Methodology,' details the experimental procedures employed. Section IV, 'Results and Discussion,' examines the data derived from these experiments. Finally, Section V, 'Conclusion,' summarizes the findings and implications of the research.

2. Theoretical Considerations

This section provides a comprehensive overview of the theoretical concepts that supports this study's approach, focusing on the evolution from traditional crack detection methods to advanced CNN architectures and the role of transfer learning in optimizing model accuracy and efficiency.

2.1. Crack Detection and Segmentation

The main challenge in crack image classification is the inhomogeneity of cracks and complex backgrounds, which often includes low contrast, shadows, uneven surfaces, and noise. Earlier methods using support vector machines (SVM) had an F1 score of 0.7359, while the introduction of deep convolutional neural networks (CNN) significantly improved performance. A 2016 study using a simple ConvNet architecture with four convolutional layers, max-pooling, a fully connected layer, and ReLU activation achieved an F1 score of 0.8965, surpassing the SVM benchmark [16]. Recent studies employed pre-trained models like VGG-19, which incorporates deeper structures to classify cracks into types such as alligator or longitudinal types, achieving an F1 score of 0.9076 [17]. Improvements in efficiency led to the use of AlexNet, which uses ReLU, dropout layers, and overlap pooling in a five-layer convolutional architecture. AlexNet along with GoogleNet, and VGG 19, achieved an F1 score of 0.99, but AlexNet had fewer layers, making it ideal for high-accuracy applications requiring lower computational loads [18].

Pixel-level crack segmentation provides detailed information on crack type, location, and severity, though challenges like pixel imbalance and down-sampling persist. Modifications to CNNs—such as removing pooling layers, using up-sampling, and adapting loss functions—help address these issues [19]. U-Net, originally developed for biomedical segmentation, has proven effective for crack segmentation [20]. In comparisons among deep CNN architectures—such as Fully Convolutional Network (FCN), Global Convolutional Network (FGN), Pyramid Scene Parsing Network (PSPNet), UPerNet, and DeepLabv3+—DeepLabv3+ achieved the highest F1 score of 0.7732, demonstrating resilience even with image blemishes [21]. A 2023 study validated YOLOv4's suitability for crack detection in complex environments, achieving 92% accuracy with 0.22 mm precision in drone-captured images, highlighting its promise for real-time applications [22]. Given these challenges, CNNs have shown considerable promise in enhancing crack detection accuracy by identifying intricate patterns within complex backgrounds. However, for tasks requiring precise localization, 2-stage CNNs such as Faster R-CNN and Mask R-CNN provide enhanced performance by isolating areas of interest with greater accuracy.

2.2. Two-Stage Convolutional Neural Network in Crack Detection

Two-Stage CNNs, including Faster R-CNN and Mask R-CNN, have advanced image detection and segmentation tasks by dividing the process into two phases: a region proposal stage for identifying areas of interest, followed by a classification and refinement stage for accurate detection. This separation allows for higher precision and adaptability in complex environments, making two-Stage CNNs particularly suitable for crack detection, where backgrounds are often heterogeneous and noisy. Traditional CNNs, while effective for image classification, face significant limitations in complex applications like crack detection. They lack mechanisms for detailed localization and segmentation, which are essential for assessing cracks' size, type, and severity. In contrast, two-Stage CNNs overcome these limitations through:

- *Enhanced Localization and Precision*: The Region Proposal Network (RPN) in 2-Stage CNNs isolates relevant regions within the image, filtering out background noise, which improves model focus on specific crack areas. This capability is crucial in infrastructure applications, where accurate localization is essential for analyzing small defects [23, 24].
- *Detailed Segmentation with Mask R-CNN*: For pixel-level segmentation, Mask R-CNN adds a segmentation branch to Faster R-CNN, allowing the model to delineate each pixel within detected cracks. This enables detailed information on crack boundaries and shapes, surpassing the capabilities of traditional CNNs in fine-grained segmentation [25]. The addition of segmentation in Mask R-CNN has been shown to improve crack localization in challenging environments [13].
- *Improved Small Object Detection*: Cracks are often small, elongated features that standard CNNs may overlook. The two-phase process in 2-Stage CNNs allows for region refinement, which enables precise detection of small details. Studies comparing CNN models, like the work by Zhang et al. (2016) [26], indicate that 2-Stage CNNs like Faster R-CNN perform significantly better in capturing fine details in infrastructure applications than single-stage models.
- *Adaptability with Transfer Learning*: Leveraging Transfer Learning, 2-Stage CNNs can use pre-trained weights from large datasets such as ImageNet to generalize on limited labeled data, enhancing both training speed and accuracy. For example, AlexNet and ResNet50 have been widely used in Transfer Learning to improve CNN generalization in crack detection, providing efficiency in limited-data scenarios like UAV-based inspections [27].

While 2-stage CNNs deliver high precision in segmentation tasks, R-CNN is found to demand substantial computational resources and operate slowly, making it less ideal for real-time applications. Mask R-CNN, likewise, requires precise parameter tuning and has a slower processing speed, limiting its utility in fast-paced environments like UAV-based monitoring. The need for a balance between speed and accuracy in UAV-based crack detection led to the selection of AlexNet for efficient classification and YOLOv4 for rapid segmentation in this hybrid approach.

2.2.1. AlexNet

AlexNet is an 8-layer CNN classifier pre-trained on the ImageNet database, capable of classifying up to 1000 object categories. The architecture includes two initial convolution layers with ReLU and max pooling, followed by three convolution layers with ReLU, max pooling, two fully connected layers with ReLU, and a final softmax classification layer. The model's efficiency and adaptability make it particularly useful in applications where lightweight, fast classification is necessary, as in UAV-based crack detection [14].

2.2.2. YOLOv4 Model

YOLOv4 is a CNN-based single-stage detection model optimized for segmentation and designed to perform real-time object detection. It uses CSPDarknet53 as a backbone, connecting to the final head with a "neck" structure, for predicting object classes and bounding boxes. Unlike 2-Stage CNNs, YOLOv4 is streamlined to operate at high speeds, making it suitable for fast segmentation in UAV-based applications. Although it lacks the dual-stage precision of models like Faster R-CNN, YOLOv4 is ideal for real-time segmentation where high speed and computational efficiency are critical [26].

2.2.3. Integration of AlexNet and YOLOv4

In this study, AlexNet is used for initial crack classification, leveraging its efficient structure to achieve high accuracy without heavy computational demands. By classifying images in this initial phase, AlexNet reduces noise, focusing YOLOv4 on relevant regions for segmentation. This combined approach optimizes both processing speed and segmentation accuracy. The hybrid model effectively balances efficiency and precision in real-time crack detection, suitable for UAV-based monitoring [27]. Table 1 outlines the different models and their respective strengths and applications.

Table 1. Two-Stage and Traditional CNN Models

Model	Type	Strengths	Weaknesses	Suitable Applications
Faster R-CNN	2-Stage	High localization, region focus	High computational demand; slower, limiting real-time use	Offline crack detection, detailed segmentation
Mask R-CNN	2-Stage	Pixel-level segmentation	Sensitive to parameter tuning; slower than YOLO	Precision tasks, boundary detection
YOLOv4	Single-Stage	Real-time segmentation	Limited with tiny, intricate cracks	Real-time crack detection, UAV-based applications
AlexNet	Single-Stage	Lightweight, fast training	Limited segmentation capabilities	High-accuracy applications with lighter load demands

Through 2-Stage CNNs and the hybrid use of AlexNet with YOLOv4, the model leverages the strengths of each approach, overcoming traditional CNN limitations in real-time crack detection for infrastructure monitoring. To further enhance the model's adaptability and performance, particularly given the limitations of data collection in UAV applications, Transfer Learning allows the use of pre-trained models to achieve robust detection with smaller datasets, reducing computational demands and enhancing accuracy.

2.3. Transfer Learning

Transfer Learning enhances CNNs by allowing models to leverage features learned from pre-trained networks on large datasets, such as ImageNet, and apply them to related tasks with limited labeled data. In traditional CNNs, training from scratch requires vast, domain-specific datasets, which are often impractical to obtain for specialized applications like UAV-based crack detection. By reusing essential feature layers trained on generic visual patterns, Transfer Learning reduces data requirements and accelerates training [28] while preserving high accuracy [29]. For example, pre-trained models like AlexNet and ResNet50 enhance generalization in crack detection, achieving faster training convergence and accuracy improvements of up to 10% over models trained solely on smaller datasets [30].

Transfer Learning improves CNN adaptability by enabling fine-tuning of higher layers specific to the target task. This approach enhances model precision in recognizing crack-related features such as irregular shapes and small-scale details. Studies show that models fine-tuned through Transfer Learning, such as Mask R-CNN, perform well in high-precision tasks, achieving F1 scores of up to 85% in complex environments [22]. Additionally, by reducing computational demands, Transfer Learning makes real-time applications feasible, supporting crack detection in GNSS-denied environments. Integrating Transfer Learning into the study's hybrid model with AlexNet and YOLOv4 optimizes both detection accuracy and efficiency for infrastructure monitoring applications [21].

2.4. Field of View

Camera calibration is performed to correct image distortion and determine the camera parameters. One of the most important intrinsic parameters is the focal length, which is the distance from the lens to the camera sensor. Camera calibration also identifies extrinsic parameters, such as distance of the camera lens from the object being captured. This information is necessary to compute the field of view (FOV) of the camera, which dictates the area a camera can capture. This is particularly essential for coverage applications like wall inspections [31]. Figure 1 illustrates the relationship between focal length, distanced, with the FOV.

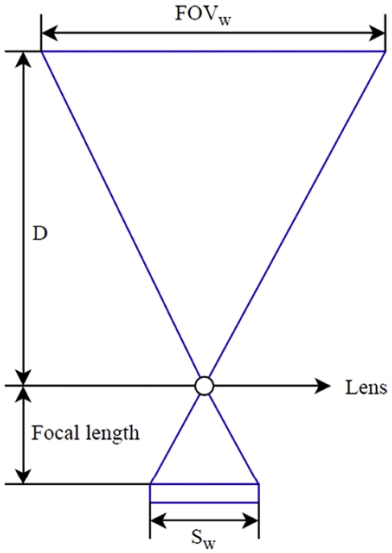


Figure 1. Focal length and field-of-view, adopted from [31]

The size of the FOV can be calculated if the distance of the camera to the object, focal length, and size of the camera sensor are determined. The computation can be found on Equations 1 and 2 with S_w and S_h being the width and height of the camera sensor, D being the distance of camera to the object, and F being the focal length.

$$FOV_w = \frac{S_w \times D}{F} \quad (1)$$

$$FOV_h = \frac{S_h \times D}{F} \quad (2)$$

Once the FOV is determined, the whole inspection area can be segmented into a grid composed of fixed-size rectangles based on the field of view's measurement. Additionally, the viewpoint's location would be the location of the lens, which is in the middle of the field of view and the proximity being the distance of the lens to the object.

3. Methods

This section outlines the methodological steps undertaken for the conduct of this experiment as shown in Figure 2 below. The study commences with identifying the materials used for the physical experimentation, followed by the model development and evaluation. Finally, the calibration and physical experimentation discusses how the model was deployed through UAV operation in a controlled environment as well as the evaluation.

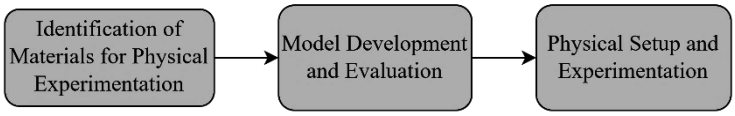
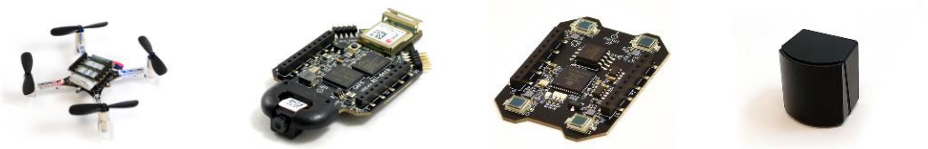


Figure 1. Research Methodology

3.1. Materials

The Crazyflie, as depicted in Figure 3, is a compact, lightweight, yet robust UAV developed by Bitcraze. Its small size and modular design make it particularly suitable for drone research, as highlighted by Chu et al. (2022) [5] for its modularity, safety, and open-source framework. This study also employs the Lighthouse Positioning Deck and the Basestation, as illustrated in Figures 3c and 3d, respectively, for drone localization and navigation.

Additionally, the AI Deck from Bitcraze is a camera module for the Crazyflie 2.X nano quadcopter, enhancing the drone with artificial intelligence capabilities. It integrates the HIMAX HM01B0 low-power monochrome camera, which has a resolution of 320×320 pixels, and a powerful GAP8 RISC-V processor, enabling real-time image processing and neural network computations for AI tasks. This modular platform supports a wide range of applications including autonomous navigation, object tracking, and environmental data collection, thereby facilitating research and development in drone-based AI technologies.



(a) (b) (c) (d)

Figure 2. (a) Crazyflie 2.1 [32], (b) AI Deck Module [33], (c) Lighthouse Positioning Deck [34], (d) Lighthouse Basestation [35]

3.2. Dataset

The dataset used in this study was gathered and published by Özgenel et al. (2018) [36]. It consists of 458 high-resolution images, resulting in 20,000 crack images of various orientations and sizes, along with 20,000 non-crack images for noise. This dataset, widely used in road crack detection studies, was split into training, testing, and validation sets with a 60:20:20 ratio. For segmentation, no publicly available dataset with bounding boxes on cracks was found. Therefore, a custom dataset was manually created by capturing crack images. To align with the classification dataset, twelve images from Özgenel's dataset were printed and pasted on a wall in various orientations; additional images were also combined to create varied types of cracks.

Using the specified camera, images of these cracks were captured at the optimal distance determined by the calibration sequence for classification and segmentation. From this data collection, 704 crack images of varying orientation and positions were obtained. These images were then transferred to a computer for manual annotation. The segmentation annotation for the involved placing a bounding box that tightly encloses each crack, minimizing unnecessary surrounding space. Similar to the classification dataset, the annotated segmentation dataset was split into training, testing, and validation sets with a 60:20:20 ratio for transfer learning.

3.3. Model Development and Evaluation

Once the datasets have been obtained, training of the classification network and segmentation network was conducted through transfer learning in MATLAB. For the classification network, a pre-trained AlexNet was used, and retrained with the Özgenel crack dataset. To match the number of classes, the last three layers of AlexNet were replaced, enabling the final classification layer to categorize images into two classes: Positive and Negative, depending on whether cracks are present. The training process involved running the network on the test dataset and adjusting it to achieve higher accuracy in subsequent iterations. Training concluded either once the maximum number of iterations was reached or when no further significant improvements in accuracy were observed. For YOLOv4, the transfer learning was also applied, and a pre-trained tiny-yolov4-coco network was used and retrained with the annotated dataset of 704 images. The model was specified to detect only a single class—the crack image. The input size was matched to the resolution of the drone's camera, and the number of anchor boxes was estimated based on the training data. Training was performed on a Ryzen 5 5500 processor and an RTX 3060 12gb graphics card.

Upon completing the training of both the AlexNet and YOLOv4 models, a two-staged CNN model was implemented in MATLAB 2023a to detect and segment cracks in structural walls. Figure 4 illustrates the proposed two-stage model: the first stage employs a crack classification model that filters input images to identify those containing cracks, while the second stage involves a crack segmentation model that delineates the cracks by placing bounding boxes around them. This methodology leverages on a combination of Alexnet and YOLOv4, which, according to the literature, has not been previously utilized in two-stage CNN systems for classification and segmentation. To evaluate the efficacy of this two-stage network, its performance was assessed using a confusion matrix and benchmarked against the standalone performance of YOLOv4, to determine whether the preliminary classification stage enhances the overall accuracy of crack segmentation.

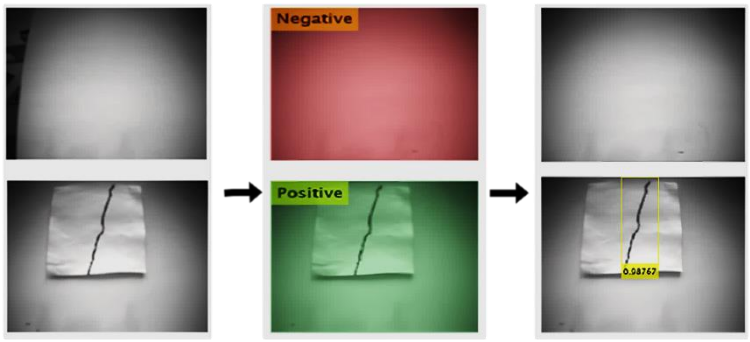


Figure 3. Proposed 2-Stage CNN Network for Classification and Segmentation

Finally, Figure 5 visualizes the output, where the resulting masks of the segmentation network are concatenated to represent the inspected area. The white regions highlight the cracks by bounding boxes which helps visualize their approximate locations. Another feature of the neural network's output is the ability to quantify the crack size. Since the field of view size is known, the pixel-to-centimeter ratio can be calculated, allowing the bounding box to serve as a scale for determining the crack's length and width. The dimensions of the bounding boxes can be converted into centimeters to determine the length and width of the crack as well as its position in the plane.

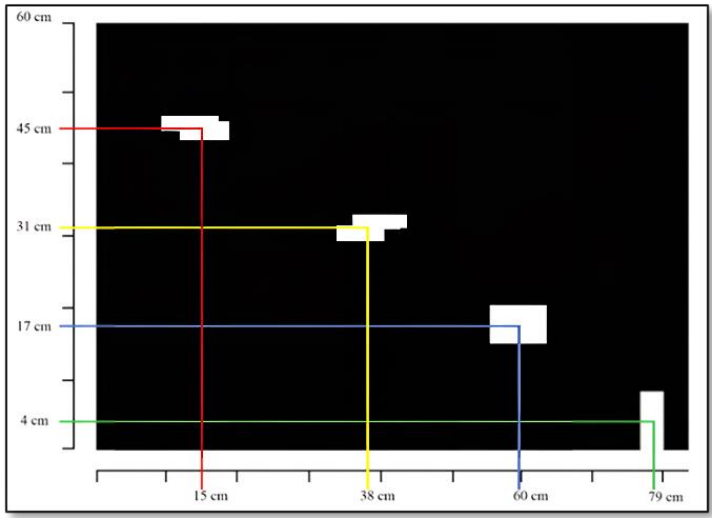


Figure 4. Crack Quantification based on Mask Size

3.4. Physical Experimentation

In this study, the imaging and localization system comprising a camera module on the AI Deck and a Lighthouse positioning deck were mounted on the UAV to gather video feedback and localize its position in the experimental space, as illustrated in Figure 6. To ensure accurate and effective data collection, the optimal distance for image capture was determined through calibration using a printed crack image and a calibration image—a chessboard with known grid sizes—positioned within the UAV's inspection area. The use of a chessboard for camera calibration is well documented for its effectiveness in determining intrinsic and extrinsic camera parameters. During the experiments, the UAV was instructed to hover and capture images at various distances (10, 13, 17, and 20 centimeters) from both the crack and chessboard images. The primary objective was to ascertain the UAV's imaging performance across these distances, pinpoint the optimal distance for effective crack classification and segmentation, and compute the camera's focal length at this distance to derive the horizontal and vertical fields of view.

The determination of the optimal distance for crack classification involved processing the captured images through the classification and segmentation network, with the system accuracy evaluated based on the proportion of images correctly classified and segmented. The furthest distance with the highest accuracy maintained was designated as the optimal imaging distance. Subsequently, an image of the chessboard captured at this optimal distance was analyzed using MATLAB to determine the camera parameters and correct any image distortions [37]. The chessboard tile dimensions facilitated estimation of the field of view, which was crucial for configuring the bounding box parameters. Throughout these calibration trials, the UAV's crack detection capabilities and corresponding accuracy levels were meticulously recorded, providing a comprehensive evaluation of the system's operational efficacy upon deployment.

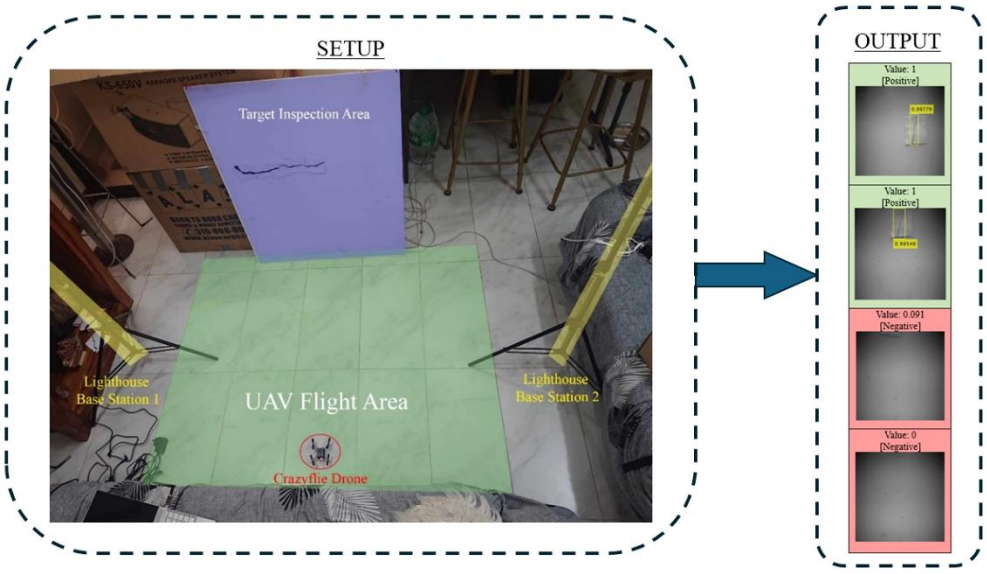


Figure 5. Experimental Setup

4. Results and Discussions

4.1. Performance Validation of Classification and Segmentation Networks

The transfer learning for AlexNet reached completion after the maximum epoch limit, with training taking 154 minutes and 42 seconds and resulting in a final validation accuracy of 99.63%. Testing on the reserved test dataset yielded an accuracy of 99.42%, closely aligning with prior studies using AlexNet on similar datasets, which reported an accuracy of 99% [15]. This consistency with previous results indicates that the retraining was effective, and that transfer learning successfully adapted the AlexNet model to the current dataset.

Similarly, the transfer learning for the YOLOv4 concluded after reaching the maximum epochs, with a total training duration of 36 minutes and 10 seconds. Upon evaluation on the test dataset, YOLOv4 achieved an average precision of 98%, which aligns with findings in related studies using YOLOv4 for crack segmentation tasks, achieving comparable precision scores [18]. This high level of precision also suggests that the transfer learning process was not only effective but also validated the model’s adaptability for segmentation in complex environments.

Across fifteen trials, the combined two-stage classification and segmentation network demonstrated consistent performance, validating the integration of AlexNet and YOLOv4 as a robust approach for crack detection tasks. These results reinforce the model’s capacity to perform well in scenarios requiring precise crack localization and segmentation, suggesting that the retrained model is well-suited for real-world applications where accuracy and adaptability are essential. The high performance also highlights the advantage of employing transfer learning in two-stage networks for achieving both efficiency and precision in infrastructure monitoring.

4.2. Camera Calibration for Enhanced Detection Accuracy in UAV Operations

Calibration is essential for the Crazyflie's crack detection capabilities, particularly given drones' limited operational times due to their compact size and battery constraints. Figure 7 illustrates a sample from the calibration experiments, with the summary of findings presented in Table 2. Images of printed cracks were captured at varying distances to assess the impact of crack-pixel ratio on detection accuracy. Given that the printed crack images were used for training, images closer to the camera, covering more field of view, achieved a higher crack pixel ratio of 2.8%, closely aligning with training conditions. At a distance of 10 cm, the neural network reached a classification accuracy of 100%, as validated by repeated trials. While the crack pixel ratio did not precisely match 2.8% at greater distances, the network maintained reliable performance.

However, testing revealed that 10 cm was impractical for Crazyflie operation, as proximity to the wall increased the risk of collision due to minor instability. Thus, the second-best distance of 13 cm was selected as the optimal distance, balancing accuracy with practical operating distance. At this distance, the field of view (FOV) was calculated as 20.5 cm x 14.9 cm, based on the chessboard calibration tile size of 0.82 cm, providing a reference for path planning. Notably, the FOV limitation of 67.9° (diagonal) likely contributed to challenges in achieving a consistent crack-pixel ratio across varied distances. This sequence is critical for enabling the Crazyflie to detect cracks efficiently, ensuring accurate detection while maintaining an operational distance that optimally balances detection accuracy with flight time and safety.

(a) (b) (c)

Figure 6. (a) Theoretical Setup of UAV hovering 13 cm from the crack image, (b) Actual Deployment of UAV hovering 13 cm from the crack image, (c) Perspective of UAV vision system and detection of a crack

Table 2. Results from Crack Detection and Segmentation

Distance from wall	Average Accuracy	Bounding Box Size
20 cm	39.50%	34.6 × 25.8 cm
17 cm	83.33%	28.8 × 21.8 cm
13 cm	88.89%	20.5 × 14.9 cm
10 cm	100.00%	17.9 × 13.4 cm

4.3. Performance of Developed Model with UAV Experimentations

Table 3 presents the confusion matrix for the classification network tested on 4,945 points during flight. The model achieved an overall accuracy of 91.5% with a precision of 84.05%, demonstrating robust performance in correctly identifying cracks. Out of the total test points, 1,671 were correctly classified as cracks (true positives), representing 33.79% of the dataset, while 2,854 points (57.71%) were accurately identified as non-crack areas (true negatives). The model achieved a recall of 94.2%, indicating its strong ability to detect the majority of actual cracks, which is essential for reliable structural assessments. Furthermore, the F1-score was calculated at 88.7%, underscoring a good balance between precision and recall and indicating that the model effectively manages both correct detections and the minimization of false positives. Notably, this accuracy aligns closely with the 92% accuracy reported by YOLOv4 in previous studies, such as those in previous studies [18, 19], showing that the two-stage CNN model leveraging transfer learning achieves comparable performance to established one-stage models.

However, the model encountered some challenges with false positives and false negatives. There were 317 false positives, accounting for approximately 6.41% of the test points, primarily due to the model mistaking edges of foam boards for cracks. Additionally, 103 false negatives (2.08%) occurred, where actual cracks near the image boundary were not detected, potentially due to lower contrast or subtle features. Despite these limitations, the model's high accuracy, coupled with its high recall and F1-score, suggests that it can reliably differentiate crack from non-crack areas in real-time UAV inspections, making it a practical solution for infrastructure monitoring.

Table 3. Confusion Matrix for Crack Detection

	Prediction Negative	**Prediction Positive**	
Actual Positive	False Negative 103 (2.08%)	True Positive 1671 (33.79%)	***Recall*** 94.20%
Actual Negative	True Negative 2854 (57.71%)	False Positive 317 (6.41%)	
		Precision 84.05%	***Accuracy*** 91.50%

Table 4 summarizes the segmentation network's performance, presented as a confusion matrix, on images classified as "containing cracks" by the AlexNet classifier. Across twelve tests and 70 identified viewpoints, the segmentation network accurately segmented cracks in 85.71% (60 out of 70 positive samples) of the images. A key advantage of the two-stage architecture was the elimination of false positives in the segmentation phase, as the classification network effectively filtered out non-crack images. Additionally, 5.71% of the images were correctly classified as true negatives, where the segmentation network ignored images that the classifier had incorrectly classified as containing cracks. However, 8.57% of images yielded false negatives, where the segmentation network missed cracks that the classifier had identified.

The overall performance metrics reveal that the segmentation network achieved a high precision of 100% due to the absence of false positives, emphasizing its reliability in avoiding unnecessary crack detections. Additionally, the recall rate of 90.9% shows the model's ability to identify most of the true cracks, a critical factor in ensuring safety in infrastructure monitoring. The F1 score, calculated at 95.2%, reflects a strong balance between precision and recall, further validating the robustness of this two-stage CNN model for crack detection. Although the model is expected to perform robustly in noisy environments, limitations in image quality—specifically due to the Himax HM01B0 camera's 324 x 324-pixel grayscale resolution—contributed to some false negatives, particularly in areas with subtle features or low contrast, where the model occasionally failed to detect visible cracks.

Despite these limitations, the model performed satisfactorily given the camera's resolution constraints, achieving an overall accuracy of 91.42%, which affirms its effectiveness in real-time crack detection under controlled testing conditions. While the model's precision is exceptionally high, suggesting robustness in avoiding false positives, reducing false negatives remains an area for further improvement. Implementing a higher-resolution camera or additional pre-processing steps may enhance the model's sensitivity to finer crack details, minimizing missed detections and broadening its applicability for real-world infrastructure monitoring.

Table 4. Confusion Matrix for Crack Segmentation

	Prediction Negative	**Prediction Positive**	
Actual Positive	False Negative 6 (8.57%)	True Positive 60 (85.71%)	***Recall*** 90.90%
Actual Negative	True Negative 4 (5.71%)	False Positive 0 (0%)	
		Precision 100%	***Accuracy*** 91.42%

4.4. Discussion and Comparative Analysis

The findings of this study highlight the efficacy of two-stage CNN models, bolstered by transfer learning, for real-time crack detection under UAV operational constraints. Achieving 91.5% accuracy in classification and 91.42% in segmentation, this model aligns closely with benchmark studies, affirming its high accuracy and demonstrating its viability for practical UAV-based applications. For example, a similar two-stage transfer learning-based approach in [7] achieved a mean Pixel Accuracy (mPA) of 92.7% and an Intersection over Union (IoU) of 88.3% for dam crack detection. Likewise, Philip et al. (2023) [9] showcased the success of ResNet50 in crack detection with transfer learning, mirroring the effective performance of AlexNet in this study. While Zhang et al. (2016) [26] leveraged a CNN-based classifier and a transformer-based network (CTv2), achieving mean F1-scores of 82.00%, 94.69%, and 92.23% on the CrackSD, CFD, and CrackSC datasets, respectively, it highlights the model's scalability and accuracy for real-world applications. Additionally, the review by Hamishebahar et al. (2022) [11] offers a comprehensive overview of crack detection techniques, situating this study within a practical spectrum that meets the real-time requirements of UAV applications. This comparative analysis, as illustrated in Table 5, supports the model's relevance and robustness, demonstrating its ability to leverage recent advancements for enhanced infrastructure monitoring despite challenges like limited camera resolution and environmental noise.

Table 5. Model Analysis for Crack Detection

Study	Model Type	Key Strengths	Limitations	Accuracy / Key Metrics	Application Context
Current Study	Two-stage (AlexNet + YOLOv4 with Transfer Learning)	High precision, adaptable to UAV-based real-time inspections	False positives due to foam edges, camera resolution constraints	Classification: F1-Score: 88.7% Accuracy: 91.5%, Segmentation: F1-Score: 95.2% Accuracy: 91.42%	UAV-based crack detection
Li et al. (2024) [7]	Two-stage (ResNet50 with SENet)	High accuracy in complex environments, improved mPA and IoU with attention mechanisms	Computationally intensive for real-time applications	mPA: 92.7%, IoU: 88.3%	Concrete dam surface crack detection
Philip et al. (2023) [9]	Transfer Learning (ResNet50, VGG16, MobileNet)	ResNet50 shown to outperform other models, strong generalization	Limitations in real-time adaptability	ResNet50 Accuracy: ~99%	Crack detection in concrete walls
Hamishebahar et al. (2022) [11]	Review (Multiple Models, including CNNs and Transfer Learning)	Comprehensive review of deep learning methods, highlights adaptability for varied applications	General limitations in CNNs for real-time crack detection	Varied depending on model	Infrastructure monitoring, multiple applications
Guo et al. (2024) [14]	Two-stage (CNN-based Classification + CTv2)	High accuracy with pixel-level precision; effective for large-scale pavement inspection	Potential limitations in detecting fine crack details; performance could vary with environmental factors	CrackSD: F1=82%, CFD: F1=94.69%, CrackSC: F1=92.23%	Pavement surface crack detection

The performance results detailed in Section 4.3 further validate the proposed two-stage CNN model. Utilizing AlexNet for classification and YOLOv4 for segmentation, the model achieved 91.5% classification accuracy and a recall rate of 94.2%, comparable to the 92% accuracy reported by YOLOv4 for crack detection applications. This alignment underscores the model's real-world applicability in UAV-based systems, where efficiency and computational economy are critical.

Transfer learning, central to this study, contributed substantial improvements over traditional two-stage CNNs. By utilizing pre-trained AlexNet and YOLOv4 models, training times were significantly reduced—154 minutes for AlexNet and 36 minutes for YOLOv4—representing a reduction of approximately 50% to 70% compared to models trained from scratch. This efficiency is particularly beneficial for UAV applications, where quick adaptation and retraining are essential. Transfer learning also enabled robust generalization with limited data, a critical advantage in structural monitoring where labeled crack images are often scarce. The model achieved high classification accuracy (99.42% on AlexNet) and segmentation precision (91.42% for YOLOv4) without extensive tuning or large datasets, effectively meeting or exceeding common benchmarks. Pre-trained layers provided foundational features, allowing consistent crack recognition under diverse conditions, which is essential for UAV deployment in varied inspection environments.

The model's precision and efficiency in real-time crack detection hold significant implications for proactive infrastructure monitoring. Early detection of minor structural defects supports timely interventions, reducing the risk of severe failures. Its adaptability to UAV-based inspections, particularly in GNSS-denied or challenging environments, enhances its utility for maintaining public safety and structural integrity. Furthermore, the model's robustness suggests potential applications beyond infrastructure monitoring. By leveraging two-stage CNNs and transfer learning, this approach offers applications in fields like urban planning, disaster response, and civil engineering. Its adaptability to real-time demands aligns with a vision for automated, continuous infrastructure assessment, where UAVs equipped with advanced detection models contribute to resilient urban systems through regular, efficient aerial inspections.

5. Conclusion

This research successfully developed and validated a two-stage convolutional neural network (CNN) model that integrates transfer learning, utilizing a novel combination of AlexNet and YOLO models for the non-destructive detection of structural cracks. Emphasizing the critical importance of timely crack detection, this approach aims to significantly enhance early detection capabilities, crucial for the proactive maintenance and safety of buildings. The model demonstrated satisfactory efficacy, achieving a classification accuracy above 90% and successfully segmenting cracks in 85.71% of the images. Additionally, the performance of the developed model was benchmarked against the results from a similar study, establishing its reliability and effectiveness. The outcomes from both simulated environments and real-world deployments confirm the model's robust capability in detecting and segmenting structural cracks, thereby reinforcing its potential as a valuable tool in structural health monitoring. These results not only affirm the model's performance but also advance the application of sophisticated machine learning techniques in the field of civil engineering. The next step of this study is to investigate a process to localize concrete cracks on wall for GPS denied environments with the use of the developed model.

6. Declarations

6.1. Author Contributions

Conceptualization, J.S. and A.Y.C.; methodology, J.S. and T.S.C.C.; software, J.S.; formal analysis, J.S. and T.S.C.C.; investigation, J.S.; resources, T.S.C.C. and A.Y.C.; writing—original draft preparation, J.S. and T.S.C.C.; writing—review and editing, T.S.C.C. and A.Y.C. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding and Acknowledgements

The proponents of this research would like to extend their gratitude to the Department of Science and Technology – Engineering Research and Development for Technology (DOST-ERDT) for providing the funds and resources necessary to conduct the study.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Lyasheva, S., Tregubov, V., & Shleymovich, M. (2019). Detection and recognition of pavement cracks based on computer vision technology. 2019 International Conference on Industrial Engineering, Applications and Manufacturing, ICIEAM 2019, 1–5. doi:10.1109/ICIEAM.2019.8742778.

[2] Kardovskyi, Y., & Moon, S. (2021). Artificial intelligence quality inspection of steel bars installation by integrating mask R-CNN and stereo vision. Automation in Construction, 130, 103850. doi:10.1016/j.autcon.2021.103850.

[3] Deng, J., Singh, A., Zhou, Y., Lu, Y., & Lee, V. C. S. (2022). Review on computer vision-based crack detection and quantification methodologies for civil structures. Construction and Building Materials, 356, 129238. doi:10.1016/j.conbuildmat.2022.129238.

[4] Kassas, Z. M., Khalife, J., Abdallah, A. A., & Lee, C. (2022). I Am Not Afraid of the GPS Jammer: Resilient Navigation Via Signals of Opportunity in GPS-Denied Environments. IEEE Aerospace and Electronic Systems Magazine, 37(7), 4–19. doi:10.1109/MAES.2022.3154110.

[5] Chu, T. S., Chua, A., Say, M., Sybingco, E., & Roque, M. A. (2022). Swarm UAV Implementation Using Radio Localization on GPS Denied Areas. ASEAN Engineering Journal, 12(3), 111–116. doi:10.11113/aej.V12.17841.

[6] Dinh, T. H., Ha, Q. P., & La, H. M. (2016). Computer vision-based method for concrete crack detection. 2016 14th International Conference on Control, Automation, Robotics and Vision, ICARCV 2016, 1–6. doi:10.1109/ICARCV.2016.7838682.

[7] Li, Q., Xu, X., Guan, J., & Yang, H. (2024). The Improvement of Faster-RCNN Crack Recognition Model and Parameters Based on Attention Mechanism. Symmetry, 16(8), 1027. doi:10.3390/sym16081027.

[8] Paramanandham, N., Rajendiran, K., Poovathy J, F. G., Premanand, Y. S., Mallichetty, S. R., & Kumar, P. (2023). Pixel Intensity Resemblance Measurement and Deep Learning Based Computer Vision Model for Crack Detection and Analysis. Sensors, 23(6), 2954. doi:10.3390/s23062954.

[9] Philip, R. E., Andrushia, A. D., Nammalvar, A., Gurupatham, B. G. A., & Roy, K. (2023). A Comparative Study on Crack Detection in Concrete Walls Using Transfer Learning Techniques. Journal of Composites Science, 7(4), 169. doi:10.3390/jcs7040169.

[10] Quintana, M., Torres, J., & Menéndez, J. M. (2016). A simplified computer vision system for road surface inspection and maintenance. IEEE Transactions on Intelligent Transportation Systems, 17(3), 608–619. doi:10.1109/TITS.2015.2482222.

[11] Hamishebahar, Y., Guan, H., So, S., & Jo, J. (2022). A Comprehensive Review of Deep Learning-Based Crack Detection Approaches. Applied Sciences (Switzerland), 12(3), 1374. doi:10.3390/app12031374.

[12] Yu, S., Jia, S., & Xu, C. (2017). Convolutional neural networks for hyperspectral image classification. Neurocomputing, 219, 88–98. doi:10.1016/j.neucom.2016.09.010.

[13] Yang, F., Huo, J., Cheng, Z., Chen, H., & Shi, Y. (2024). An Improved Mask R-CNN Micro-Crack Detection Model for the Surface of Metal Structural Parts. Sensors, 24(1), 62. doi:10.3390/s24010062.

[14] Guo, F., Liu, J., Xie, Q., & Yu, H. (2024). A two-stage framework for pixel-level pavement surface crack detection. Engineering Applications of Artificial Intelligence, 133, 108312. doi:10.1016/j.engappai.2024.108312.

[15] Chen, Z., Wang, C., Wu, J., Deng, C., & Wang, Y. (2023). Deep convolutional transfer learning-based structural damage detection with domain adaptation. Applied Intelligence, 53(5), 5085–5099. doi:10.1007/s10489-022-03713-y.

[16] Brodzicki, A., Piekarski, M., Kucharski, D., Jaworek-Korjakowska, J., & Gorgon, M. (2020). Transfer learning methods as a new approach in computer vision tasks with small datasets. Foundations of Computing and Decision Sciences, 45(3), 179-193. doi:10.2478/fcds-2020-0010.

[17] Hammouch, W., Chouiekh, C., Khaissidi, G., & Mrabti, M. (2022). Crack Detection and Classification in Moroccan Pavement Using Convolutional Neural Network. Infrastructures, 7(11), 152. doi:10.3390/infrastructures7110152.

[18] Rajadurai, R. S., & Kang, S. T. (2021). Automated vision-based crack detection on concrete surfaces using deep learning. Applied Sciences (Switzerland), 11(11), 5229. doi:10.3390/app11115229.

[19] Tang, Y., Zhang, A. A., Luo, L., Wang, G., & Yang, E. (2021). Pixel-level pavement crack segmentation with encoder-decoder network. Measurement: Journal of the International Measurement Confederation, 184, 109914. doi:10.1016/j.measurement.2021.109914.

[20] Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., & Shoaib, M. A. (2022). Structural crack detection using deep convolutional neural networks. Automation in Construction, 133, 103989. doi:10.1016/j.autcon.2021.103989.

[21] Wang, J. J., Liu, Y. F., Nie, X., & Mo, Y. L. (2022). Deep convolutional neural networks for semantic segmentation of cracks. Structural Control and Health Monitoring, 29(1), 2850. doi:10.1002/stc.2850.

[22] Kao, S. P., Chang, Y. C., & Wang, F. L. (2023). Combining the YOLOv4 Deep Learning Model with UAV Imagery Processing Technology in the Extraction and Quantization of Cracks in Bridges. Sensors, 23(5), 2572. doi:10.3390/s23052572.

[23] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031.

[24] Bai, T. (2020). Analysis on Two-stage Object Detection based on Convolutional Neural Networks. Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2020, 321–325. doi:10.1109/ICBASE51474.2020.00074.

[25] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 386–397. doi:10.1109/TPAMI.2018.2844175.

[26] Zhang, L., Yang, F., Daniel Zhang, Y., & Zhu, Y. J. (2016). Road crack detection using deep convolutional neural network. Proceedings - International Conference on Image Processing, ICIP, 2016-August, 3708–3712. doi:10.1109/ICIP.2016.7533052.

[27] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into Imaging, 9(4), 611–629. doi:10.1007/s13244-018-0639-9.

[28] Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning?. arXiv preprint arXiv:1608.08614. doi:10.48550/arXiv.1608.08614.

[29] Chandrababu, G., Thomas Lee, O., & Rekha, K. S. (2022). An Abridged Review of Transfer Learning Technology. Proceedings 2nd International Conference on Next Generation Intelligent Systems, ICNGIS, 1–7. doi:10.1109/ICNGIS54955.2022.10079859.

[30] Matarneh, S., Elghaish, F., Rahimian, F. P., Abdellatef, E., & Abrishami, S. (2024). Evaluation and optimisation of pre-trained CNN models for asphalt pavement crack detection and classification. Automation in Construction, 160, 105297. doi:10.1016/j.autcon.2024.105297.

[31] Tan, Y., Li, S., Liu, H., Chen, P., & Zhou, Z. (2021). Automatic inspection data collection of building surface based on BIM and UAV. Automation in Construction, 131, 103905. doi:10.1016/j.autcon.2021.103881.

[32] McGuire, K. (2024). Crazyflie 2.1 Back in Stock!. Bitcraze AB, Malmö, Sweden. Availbale online: https://www.bitcraze.io/products/crazyflie-2-1/ (accessed on May 2024).

[33] Bitcraze. (2024). AI Deck 1.1. (Brain Boost: Artificial Intelligence in a Nutshell). Bitcraze AB, Malmö, Sweden. Availbale online: https://www.bitcraze.io/products/ai-deck/ (accessed on May 2024).

[34] Bitcraze. (2024). Lighthouse Positioning Deck. Bitcraze AB, Malmö, Sweden. https://www.bitcraze.io/products/lighthouse-positioning-deck/ (accessed on May 2024).

[35] Bitcraze. (2024). Lighthouse V2 Base Station. Bitcraze AB, Malmö, Sweden. https://store.bitcraze.io/products/lighthouse-v2-base-station (accessed on May 2024).

[36] Özgenel, Ç. F., & Sorguç, A. G. (2018). Performance comparison of pretrained convolutional neural networks on crack detection in buildings. In Isarc. proceedings of the international symposium on automation and robotics in construction, IAARC Publications, 35, 1-8. doi:10.22260/ISARC2018/0094.

[37] De la Escalera, A., & Armingol, J. (2010). Automatic chessboard detection for intrinsic and extrinsic camera parameter calibration. Sensors, 10(2), 2027-2044. doi:10.3390/s100202027.