ISSN No:-2456-2165

Attention-Based Automated Pallet Racking Damage Detection

Mujadded Al Rabbani Alif
Department of Computer Science
Huddersfield University
Huddersfield, HD1 3DH, United Kingdom

Abstract:- Pallet racking systems are shelves that are specifically intended to hold palletised items, and they are essential for the safe and effective handling of products in warehouses. These shelves are susceptible to damage from a variety of sources, including as wear and tear and collisions, which might jeopardise their structural integrity and put workers and stored items at risk. It's critical to identify faulty pallet racking quickly to avoid mishaps, product loss, and interruptions to business operations. Pallet racking system upkeep and routine inspections, however, can be expensive and prone to human mistakes. This research study suggests Pallet-Net, a unique deep learning technique that employs an attention-based convolutional neural network (CNN) to automatically detect faulty pallet racking, as a solution to this problem. The suggested technique uses attention processes to concentrate on the pallet racking image's damaged areas, making it easier to locate and identify damage. Pallet-Net precisely categorises the racking as either damaged or undamaged by learning the discriminative properties of these zones. The suggested approach, when compared to previous studies, provides great robustness and accuracy in locating and recognising damaged areas in pallet racking photos. Moreover, the proposed method obtains a 97.64% total accuracy rate, with 98% precision, 98% recall, and 98% F1 score. Recent deep learning models like Vision Transformer (ViT) and Compact Convolutional Transformer (CCT) are also analysed and compared to the suggested architecture.

Keywords:- *Pallet Racking Systems; Logistics; Material Handling; Structural Integrity; Deep Learning; Attention Mechanisms; Convolutional Neural Networks; Image Classification; Spatial Transformer Network; Vision Transformer; Compact Convolutional Transformer.*

I. INTRODUCTION

Pallet racking refers to a system of storage racks specifically designed to organise and efficiently store goods in warehouses and storage facilities. It consists of vertical frames, horizontal beams, and various supporting components to create multiple levels of storage space. It is pivotal in efficiently storing and organising goods in warehouses and storage facilities. These systems provide vertical storage solutions that maximise space utilisation and enable easy access to stored items without requiring additional floor space. By providing an organised storage solution, pallet racking allows efficient inventory control, easy product identification, and streamlined picking operations, reducing time and effort. However, pallet racking systems are susceptible to damage over time due to various factors, including collisions, overloading, improper handling, wear and tear, incorrect installation or maintenance, and external forces. Accidental collisions with forklifts or other equipment can result in bending, distortion, or misaligning of structural components, such as upright frames and horizontal beams. Exceeding the weight capacity of the racks can lead to structural strain, compromising stability and potentially causing collapse. Inadequate handling practices and improper placement or removal of pallets can exert excessive force on the racking system, resulting in impact damage. Wear and tear from continuous loading and unloading, environmental conditions, and friction can gradually weaken the rack's structural integrity, leading to rust, corrosion, or deterioration. Additionally, external forces like earthquakes, extreme weather conditions, or impacts from heavy objects can threaten the integrity of pallet racking systems [1], compromising their structural integrity and posing significant risks to personnel and stored goods. Detecting and addressing damaged pallet racking in a timely manner is essential to prevent accidents, minimise product loss, and ensure the smooth operation of warehouse logistics.

The conventional approach to identifying damaged pallet racking heavily relies on manual inspections carried out by trained personnel. During these inspections, the racking system is visually examined for indications of damage, such as bent components, cracks, or misalignments. While this technique serves as a starting point for detection, it has several shortcomings. For one, manual inspections are time-consuming and demanding, particularly in large-scale warehouses or facilities with numerous racks, causing delays and disruptions to daily operations. Secondly, the subjectivity of visual assessments introduces the possibility of human error, leading to missed or misidentified damages. The interpretation of damage severity may also vary among different individuals, further affecting the consistency of detection results. Furthermore, manual inspections may not effectively detect subtle signs of damage or potential structural weaknesses that could result in accidents or failures in the future. Additionally, these inspections offer limited quantitative data for analysis and tracking of the overall health and condition of the racking system. In summary, the manual inspection approach for detecting damaged pallet racking requires greater efficiency, consistency, and the ability to provide comprehensive insights for effective maintenance and risk management.

Recent advances in computer vision and machine perception using deep learning promise automated solutions in diverse areas, including healthcare [2,3], renewable energy [4], and industrial quality inspection [5]. In this research, we leverage such techniques to develop automated damage detection for warehouse pallet racking systems, critical but susceptible components of inventory storage. The field has witnessed dramatic progress through sophisticated deep neural network architectures such as convolutional neural networks (CNNs) [6] and recurrent neural networks (RNNs) [7], enabling unprecedented performance in computer vision, language processing, and speech recognition. Landmark CNN models, including VGGNet [8], ResNet [9], Inception [10], RCNN [11], and Fast RCNN [12], aided by expanding datasets and GPU computing, have achieved remarkable accuracy in classification, detection, and generative modelling of images. Novel deep approaches like YOLOv7 [13], GANs [14] and vision transformers [17] further extend these abilities. Building upon such advances, we propose a tailored CNN methodology employing visual attention to focus selectively on racking damage cues, learning highly discriminative representations, and enabling precise automated identification. By pursuing robust, accurate, and computationally efficient perception, this research aims to promote safety and efficiency in the automated monitoring of warehouse storage environments.

This paper puts forward Pallet-Net, a novel computational method leveraging attention-driven convolutional neural networks (CNNs) to automatically classify warehouse pallet racking systems as damaged or undamaged. Explicitly focusing visual attention on areas indicative of damage facilitates precise localisation and identification of distorted, cracked, or misaligned rack structures from images. Our tailored CNN architecture, trained on pallet rack datasets, learns to extract highly discriminative damage characteristics, enabling reliable automated decisions on rack integrity. We extensively evaluate Pallet-Net against recent methods, including vision transformers and compact convolutional transformers. Experiments demonstrate state-of-the-art classification accuracy, precision and recall exceeding 97% on held-out test data, with computational efficiency amenable to real-time monitoring. By reliably automating visual assessments currently requiring laborious manual inspection, this research promises significantly enhanced safety and reduced downtimes and risks in modern warehouse environments. Results further inform the future incorporation of attention-based deep learning in related structural health monitoring applications.

The remainder of this paper is organised as follows. Section 2 discusses related work on object detection and classification using deep learning. Section 3 presents the methodology, including dataset collection, network architectures, training process, and ensemble learning techniques. Section 4 describes the experimental setup, evaluation metrics, and results. Section 5 compares our proposed solution with other existing solutions, and Section 6 concludes the paper.

II. RELATED WORK

A. Pallet Racking Inspection Methods

Numerous studies have been conducted on detecting and evaluating damaged pallet racking systems. The conventional approach entails manual inspections by trained personnel who visually examine the racking systems for indications of damage, such as bent or distorted components, cracks, or misalignments. However, these inspections could be more laborious, time-consuming, and susceptible to human error. To overcome these limitations, various automated inspection techniques have been researched by experts. In a recent study, Hong-Hu Zhu et al. [18] delved into the increasing usage of innovative sensing technologies in civil infrastructure and their advantages in construction, operation, maintenance, and upgrading. He highlighted various facets of innovative sensing technologies and their utilisation in civil infrastructures, such as innovative mechanisms and devices, on-site implementation, supporting technologies and methodologies, and real-life examples. In another recent research paper, Hussain et al. [19] presented a self-governing system for inspecting storage racks using the MobileNetV2-SSD architecture. The proposed system is claimed to be utilised in distribution centres, warehouses, and retail store facilities, as it has a mean average precision of 92.7% and can extend its coverage to higher-level racking with the help of a forklift cage. The authors compiled the first racking dataset for this study based on actual pallet racking images from various operational warehouses.

Furthermore, they plan to improve the solution by including several damage detection classes and collaborating with SEMA to develop a defect detection architecture. Moreover, Chuan-Zi Dong et al. [20] in their research provided an overview of computer vision–structural health monitoring (CV-SHM) at local and global levels for element, crack, delamination, displacement, vibration, modal identification, load factor estimation, and structural identification. The author described CV-SHM as an excellent complement to conventional SHM due to its advantages, such as non-contact measurements, long-distance data collection, low cost, and reduced labour with minimum interference or intrusion to the daily operation of structures. Hussain et al. [21] introduced a framework centred on the YOLOv7 architecture in a different study. The framework includes a domain variance modelling mechanism to address data scarcity, resulting in a mean average precision of 91.1%. This solution offers a non-invasive approach to defect detection that differs from conventional sensor-oriented methods and can potentially reduce client costs.

B. Object Detection and Classification with DL

Deep learning (DL) has shown remarkable success in various computer vision tasks, including object detection and classification. Numerous studies have explored the application of CNNs for accurate and efficient detection and classification of objects in images.

Techniques such as Faster R-CNN, YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector) have been widely adopted for object detection. Zaidi et al. [22] published a study on object detection methods in the modern world. The study also covers contemporary lightweight classification models used on edge devices. The need for lightweight models that can be deployed on mobile and embedded systems is increasing, and the study shows how various object detectors have developed. Similarly, in their paper, Liu et al. [23] review deep learning methods for detecting small objects in images, including challenges and solutions, practical techniques, and related research areas. The paper compares the performances of leading deep learning methods, including YOLOv3, Faster R-CNN, and SSD, based on three large benchmark datasets of small objects. The experimental results show that while the detection accuracy on small things by these deep learning methods was low, Faster R-CNN performed the best, while YOLOv3 was a close second. Finally, in their research, Yang et al. [24] propose a real-time tiny-part defect detection method for manufacturing using deep learning algorithms. The authors establish a correlation model between the part system's detection capability coefficient and the conveyor's moving speed and propose a defect detection algorithm based on a single short detector network (SSD) and deep learning. The paper also addresses the problem of missed detection using an industrial real-time detection platform and a missed detection algorithm based on intermediate variables. These methods leverage CNNs to extract image features and employ region proposal mechanisms or anchor-based approaches to identify object-bound boxes.

In the context of object classification, CNN architectures like AlexNet, ResNet, and EfficientNet have been widely used. These models leverage deep convolutional layers to capture hierarchical features and achieve high classification accuracy. In their paper, Akinosho et al. [25] compare the performance of edge detection algorithms and deep convolutional neural networks (DCNN). The authors analyse a dataset of 19 concrete images and compare the relative performance of six typical edge detection schemes and the AlexNet DCNN architecture in different modes. The edge detection methods accurately detected 53-79% of cracked pixels. Still, they produced residual noise in the final binary images, whereas DCNNs accurately labelled pictures with 99% accuracy and detected much finer cracks than edge detection methods.

Similarly, Weimer et al.[26] explore the use of DCNN for defect detection in industrial inspection instead of manually engineering features. According to the author, DCNN automatically generates powerful features through hierarchical learning strategies from massive training data with minimal human interaction. The proposed approach is tested on a dataset with 12 different classification categories of visual defects occurring on heavily textured backgrounds, with excellent results and low false alarm rates. Liu et al. [27] explore the application of robots in intelligent supply chains and digital logistics to perform efficient operations, energy conservation, and emission reduction in warehousing and sorting. The researchers established an image recognition model using a convolution neural network (CNN) to identify and classify goods by simulating a human hand-grasping object.

C. Attention Mechanisms in CNNs

Attention mechanisms allow neural networks to emulate biological perception and cognition by selectively prioritising the most task-relevant visual information. This targeted focus on salient environmental cues drives enhanced efficiency and accuracy even where critical visual signatures occupy just a fraction of the full sensory space. Tang et al.'s [28] manufacturing damage classification framework first combined spatial attention with convolutional neural networks (CNNs) to achieve 93.3% accuracy, significantly improving on previous approaches lacking such selective computational focus. Follow-up research by Su et al. [29] validated complementary attention mechanisms for suppressed noise and improved solar cell defect identification. In agricultural applications, Shahi et al. [30] integrated CNN features with attention-based modules to enable automated fruit classification as the first stage of precision harvesting. Collectively, these works presage automation across tedious, inconsistent manual structural monitoring tasks spanning warehouse, manufacturing, solar, and agricultural sectors. Building upon these latest developments at the intersection of computational perception and selective focus, we propose an attention-driven CNN methodology to reliably detect hazardous pallet racking distortions in inventory storage environments. Our approach learns highly discriminative damage characteristics to match or enhance human visual assessments while integrated attention filters out task-irrelevant cues. More broadly, research into such biomimetic selectivity and efficiency gains continues to advance a new generation of intelligent systems endowed with heightened situational awareness for reliable autonomous decision support. As these technologies fundamentally disrupt sectors centred upon human evaluation, policy and regulation must proactively address emerging societal impacts.

Although several image classification techniques have been utilised to classify damaged pallet racking, deep learning methods have recently gained significant attention. Attention mechanisms can be a valuable tool to improve the accuracy and robustness of the models. Recent deep learning models such as Vision Transformer and Compact Convolutional Transformer have shown potential in improving the speed and accuracy of image classification.

III. METHODOLOGY

A. Dataset

Effective machine learning relies on comprehensive and representative datasets encompassing real-world complexity; for our pallet-racking damage detection system, labelled images depicting various distortion types are needed to train computational models and quantify evaluation generalizability. We gathered on-site photos of bent beams, cracked uprights, misalignments and other visible defects from Tile Easy and Lamteks warehouses. As public pallet racking datasets remain unavailable, this collection provides an essential bootstrap capturing noise, occlusion and variability challenging unaided human assessments. While expanding sample diversity and quantity would further enhance robustness, these initial images enable the development of a rigorous methodology that assesses authentic damage manifestations rather than simulated data.

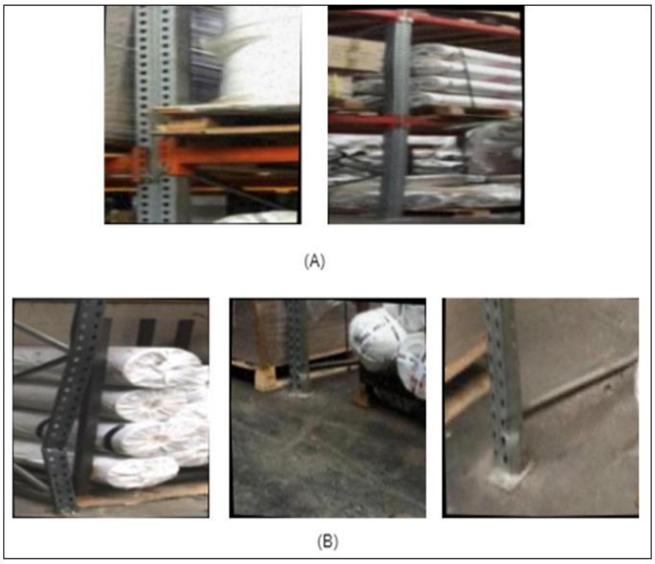

Fig 1 Pallet Racking dataset Samples (A) Normal and (B) Damaged

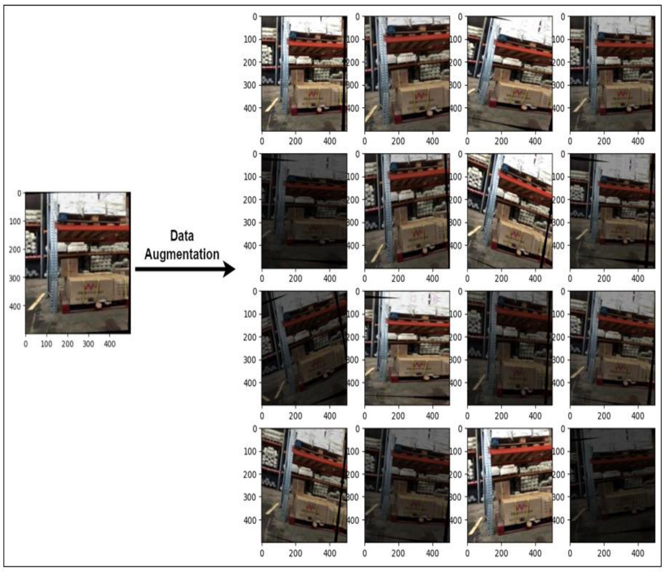

Fig 2 The Effects of data Augmentation on the Training and test Images.

Data collection leveraged an iPhone 8 12MP camera selected for sensor fidelity matching our targeted Raspberry Pi edge deployment. Images simulate views from a forklift-mounted rack inspection system, withstanding volatility from motion, occlusion and variable lighting. A human operator proxy holding the smartphone-based camera towards storage racking emulates automated on-vehicle assessments' precise positional dynamics and visual perspective. This contextual data gathering aims to furnish models with representation crucial for a smooth transition from laboratory to materials handling environments. Additionally, mobile visual data promises scalability via crowdsourcing to rapidly expand sample diversity in future work.

Figure 1 exemplifies dataset diversity across undamaged and damaged pallet racking images. Healthy warehouse storage infrastructure constitutes relatively simple classification tasks targeting racking components against static backgrounds (Fig. 1A). However, distortion severity varies extensively among damaged samples (Fig. 1B), challenging human evaluation consistency, especially for subtle cases. The centre image depicts a rack leg crack that could easily elude unaided visual assessment compared to the obvious right deformation. All samples embed environmental context, including occlusion, variable lighting and noise. Augmentation must, therefore, balance class distinction and resolution preservation with realistic domain complexity to enable effective model generalisation. Overall, these images capture the multi-scale damage phenomena, ambiguity and scene diversity demanding selective, context-aware computational focus - an ideal testbed to advance attention-based automated inspection.

By collecting and curating this initial dataset, we aim to provide a foundation for training and evaluating the proposed autonomous racking inspection mechanism using CNN models with attention. The dataset offers a diverse range of normal and damaged racking images, enabling the model to learn and generalise patterns associated with different racking conditions.

This preliminary dataset establishes an essential benchmark for developing and evaluating automated pallet-racking assessment systems using attention-focused computational perception. Despite sample size constraints, the images capture real-world diversity across damage modes and environmental variability. More broadly, benchmarking on authentic anomalies rather than simulated data should enhance model generalisation to the complexities of deployable structural monitoring.

B. *Data Augmentation*

Data augmentation enables the artificial expansion of limited training sets to enhance model generalisation - mimicking the diversity of real-world phenomena from limited samples. Popular techniques add noise or apply transformations like rotation while retaining core semantics. Such expanded sets curb overfitting, improve resilience to previously unseen inputs, and strengthen the mapping from images to damaged phenotypes learned during training. We leverage Keras' [31] flexible ImageDataGenerator toolkit, which has become a vital utility across deep learning applications owing to its simplicity and built-in transforms. Although constrained generalisation demands eventually surpass synthetic expansion alone, augmentation grants valuable bootstrapping for developing rigorous defect detection from scarce racks lacking comprehensive historical assessments.

Effective automation requires resilience across damage modes, environments and operating conditions. We augment pallet racking data with techniques including brightness adjustment, rotation, zooming and shear transformations (Fig. 2). This expanded, distorted sample diversity compels models to generalise rather than memorise, improving deployable decision-making amid complex warehouses far beyond constrained training distributions.

- *Feature-Wise Normalization:*

Input consistency is crucial for model convergence. Feature-wise normalisation (Equation 1) rescales inputs to constrain variability - transforming dimensions to standard normal distributions with zero mean and unit variance. This harmonic representation attenuates the influence of noise and distortions, so computational focus targets the underlying damage morphology rather than incidental data properties. Ultimately, learning intrinsically invariant causal markers promises improved generalisation.

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (1)$$

Where σ is the standard deviation, μ is the mean value of the feature throughout the dataset, and x is the input feature. Every feature dimension is subjected to a separate feature-wise normalisation method, guaranteeing that every feature has a mean of zero and a standard deviation of one. The model's capacity to tolerate differences in the distribution of input features is improved when feature-wise normalisation is applied during data augmentation. This method successfully reduces the effect of variations in brightness, contrast, or intensity between different photographs.

- *Feature-Wise Centre*

Another efficient technique for augmenting data in deep learning to enhance model performance and generalisation is feature-wise centring. Using this method, the appropriate input feature is subtracted from the mean value of each feature dimension. The feature-wise centring equation is defined by equation (2).

$$x_{centered} = x - \mu \quad (2)$$

Where μ is the feature's average value over the dataset, and x is the input feature. During data augmentation, the model's sensitivity to the mean value of features can be decreased by using feature-wise centring. By using this method, the impact of differences in brightness or intensity levels across samples is lessened [33]. The model can more effectively identify the relative differences and patterns linked to broken pallet racking by centring the features, which prevents the model from being impacted by overall changes in the input data.

- *Shearing*

Shearing is a widely used deep learning data augmentation technique that modifies input data geometrically. It involves skewing or tilting pictures along a certain axis to distort them. Equation (3) is an expression for the shearing transformation.

$$\begin{bmatrix} x_{sheared} \\ y_{sheared} \end{bmatrix} = \begin{bmatrix} 1 & shear_{factor} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

- *Others*

We have included several widely used data augmentation strategies in addition to the ones that were previously described.

- *Zoom:*

This augmentation entails applying a certain zoom factor to the supplied image. By using this method, the model may be trained to recognise and categorise broken pallet racking at various sizes and scales, mimicking the variances in object sizes and distances found in the real world.

- *Rotation:*

The supplied picture is transformed via rotation in this augmentation. This method improves the model's capacity to handle multiple viewing angles by helping it identify faulty pallet racking from a variety of viewpoints or orientations. Equation (4) was utilised to compute the rotation in order to facilitate this augmentation.

$$H = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \quad (4)$$

- *Brightness:*

The input image's brightness level is adjusted by this augmentation. Brightness changes strengthen the model's resistance to various lighting scenarios and guarantee correct classification even in the presence of fluctuating illumination.

- *Width And Height Shift:*

This augmentation entails a horizontal or vertical picture shift. With the use of this method, the model may be trained to identify broken pallet racking even in situations when it is only partially visible or positioned differently in the picture.

- *Fill Mode Reflects:*

This augmentation takes care of any voids or regions left by previous augmentations, such as rotation or shifting. By reflecting the adjacent pixels, it fills in the empty pixels and keeps the image whole.

By combining these methods of data augmentation, we are able to provide an enhanced dataset that depicts changes in pallet racking that have been damaged. Furthermore, the diversity and number of training samples are greatly increased by the enhanced dataset, which aids in the model's improved generalisation and classification performance [34].

Colourisation imparts limited semantic insight for structural damage classification, instead obstructing the perception of subtle depth or texture distortions with incidental hue variations. We deploy grayscale transformation, a technique shown by Li et al. [35], to improve dermatological anomaly detection models to similarly enhance rack damage cognition. Eliminating RGB colour space dimensionality focuses computations exclusively on luminance-linked cues while enabling simplified model architectures.

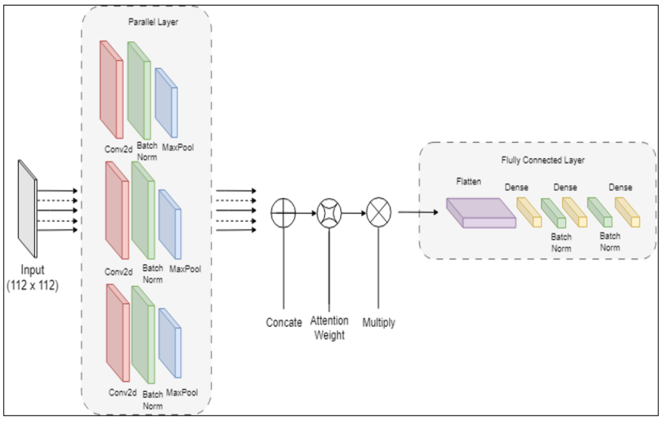

Fig 3 Pallet-Net architecture.

Grayscale's single channel mitigates inter-channel correlation, imposing ineffective representational constraints for convolutional Filter learning. Attenuating colour information steers models towards crucial shape and morphology factors rather than superficial chromatic tendencies counterproductive to generalisable decisions. Overall, this restrictive representation learning approach filters out rack imaging noise to improve accuracy - exploiting intrinsic intensity patterns correlated with damage while discarding nuisance colour variation. More broadly, task-specific dimensionality reduction that isolates primary explanatory factors epitomises efficient biological perception for accelerated anomaly cognition. This bio-inspired sparsity simultaneously enhances model performance and computational efficiency, which is crucial for embedded structural monitoring.

The damaged pallet racking categorisation work has been conducted consistently using an image size of 112x112 pixels throughout our investigation. This calculated choice was made with a number of factors in mind in an effort to increase our model's precision and effectiveness. Initially, maintaining a constant picture size guaranteed consistency in the input data fed into the model for both training and inference. Our model was able to acquire and derive significant characteristics from photographs of damaged pallet racking, regardless of the images' initial size, because of this constancy. It made comparing and analysing the various photographs in the dataset easier as well.

Additionally, the 112x112 pixel standard picture size contributed to a decrease in memory use and computational complexity [36]. Furthermore, the key characteristics and details of the damaged pallet racking were preserved because of the 112x112 picture size. It struck a compromise between lessening the computing load and collecting enough geographic information. Last but not least, this scale made sure the model could pick up on essential patterns, textures, and structural details connected to broken pallet racking without adding superfluous detail or excessive noise that might compromise categorisation accuracy.

Robust automation centres on harmonising model simplicity, efficiency and real-world performance through representation learning heuristics tailored to the application. Our image standardisation, grayscale conversion and data augmentation synergistically filter pallet-racking data complexity down to the core factors explicating damage morphology. By eliminating incidental colour variation while exposing models to an expanded, distorted sample distribution, we steered computation towards intrinsic intensity patterns predictive of actionable rack defects. Meanwhile, consistent image resizing removes confusing variability that might inhibit convolutional filter convergence. Together, these techniques significantly enhanced classification accuracy by reducing the burden of memorisation and overfitting intrinsic to limited data. More broadly, such complexity reduction through domain knowledge infusion epitomises efficient biological perception - discarding sensory noise to amplify causal signatures. Our methodology thus demonstrates how even modest datasets can fuel deployable decision automation so long as data curation targets explanatory factors using time-tested bio-inspiration.

C. Detailed Description of the CNN Architectures

As a consequence of our study, a unique CNN architecture called Pallet-Net—an integrated attention mechanism—was created with racking inspection in mind. This design efficiently separates pallet racking that is damaged from that that is not. We have also experimented with additional state-of-the-art deep learning architectures, such as Custom Compact Convolutional Transformer (CCT) and Custom Vision Transformer (VIT), based on current findings.

➢ *Pallet-Net*

The Pallet-Net's architecture consists of three CNN connections operating in parallel, each with a distinct 3x3, 5x5, and 7x7 kernel size. In Figure 3, the architecture is displayed. The input goes through batch normalisation, max-pooling, and convolutional operations in each connection. The final feature maps obtained from the three connections are concatenated, and then they are run through a dense layer using a SoftMax activation function and one unit.

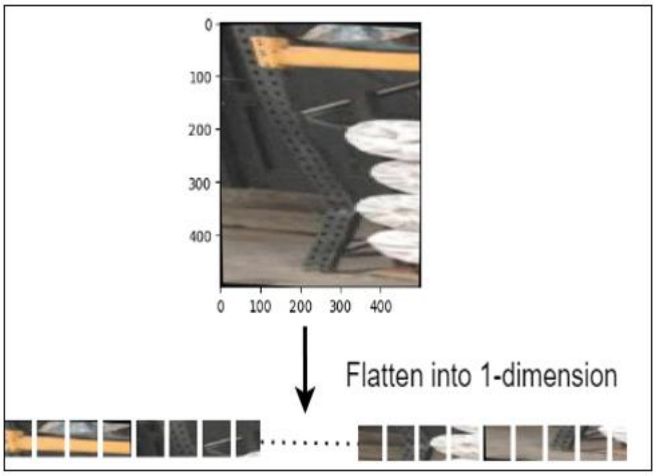

Fig 4 Image Flattening in to 1-Dimension in a Layer

Pallet-Net's concatenation technique served as the foundation for the attention mechanism, which generated attention weights for each feature map that the parallel connections created. As seen in Figure 4, the attention output that results from multiplying the attention weights by the concatenated feature maps is shaped into a 1-dimensional array via a flattened layer.

Two fully linked layers with 512 and 256 neurons each make up Pallet-Net. Each layer makes use of batch normalisation and the ReLU activation function, which was chosen for its straightforward mathematical formulation and given in Equation (5).

$$f(x) = \max(0, x) \tag{5}$$

For the purpose of automatically identifying and categorising broken pallet racking, the Pallet-Net architecture with an integrated attention mechanism is a useful method. The model performs better because the attention mechanism creates attention weights for each feature map that the parallel connections produce. The accuracy and resilience of the model are greatly enhanced by the simultaneous convolutional connections and the attention method. Here is an equation that may be used to represent the suggested model: 6,7,8,9,10,11, and 6.

$$f(x) = \max\big(C = Concatenate(P_1, P_2, P_3)\big) \tag{6}$$

$$W = Softmax(Dense(C)) \tag{7}$$

$$O = C \odot W \tag{8}$$

$$F = Flatten(O) \tag{9}$$

$$H_1 = Dense(512, ReLU)(F) \tag{10}$$

$$H_2 = Dense(256, ReLU)(H_1) \tag{11}$$

$$Y = Dense(n_{classes}, Softmax)(H_2) \tag{12}$$

In this case, C represents the concatenated output of the parallel connections, W represents the attention weights calculated using a dense layer with softmax activation, and X represents the input layer of shape (image_size, image_size, P1, P2, P3). H1 and H2 stand for the first and second completely connected layers, O for the attention output derived by element-wise multiplying C and W, F for the flattened output of O, and Y for the output layer with a class number of neurons. This formula provides a succinct mathematical depiction of the customised attention-based CNN model by symbolically representing the model's layers and processes. the following is a detailed description of pallet-net's architecture:

- *Input Layer:*

Given that the input layer's shape is (112,112,1), the model is likely to accept grayscale photos of 112 x 112 size.

- *Parallel Convolutional Layers:*

The network has three convolutional layers arranged in parallel. There are 32 filters per layer, with three, five, and seven-by-seven-inch filters in each size. All layers get the application of the ReLU activation function. Every parallel connection has a 2x2 max pooling and batch normalisation layers.

- *Concatenation:*

This node concatenates the outputs from the three parallel connections.

- *Attention Mechanism:*

Using a dense layer of 1 unit and softmax activation, we apply an attention mechanism to the concatenated output to derive attention weights. The element-wise product of the concatenated output and the attention weights yields the final attention output. Equations (13 and 14) represent the attention mechanism equation.

$$E = Softmax\left(\frac{QW_q . KW_K^T}{\sqrt{d_k}}\right) \tag{13}$$

$$C = E(VW_v) \tag{14}$$

Here, the network or decoder's current state is represented by the query vector Q; the input or encoder states are represented by the set of key vectors K; the1 input or encoder states are represented by the set of value vectors V, the importance weights assigned to the input states are represented by the attention matrix E, and the context vector C is calculated as the weighted sum of the value vectors. The scaled dot product between the query and key vectors is multiplied by the Softmax function to obtain the attention matrix E in this formula. Each input state's weight or relevance is represented in the resultant attention matrix. Next, the attention matrix is multiplied by the value to get the context vector C.

- *Flatten Layer:*

The attention output is flattened as input to thick layers in order to conform to the processing step.

- *Fully Connected Layers:*

Pallet-Net has two completely linked dense layers that come after the convolutional layers and the attention component. Relu activation is included in the first layer's 512 units and the second layer's 256 units. For performance regularisation, a batch normalisation layer is also included in each layer.

- *Output Layer:*

In order to generate the output, the final layer is dense with three units and SoftMax activation.

➢ *Custom Vision Transformer (ViT)*

For benchmarking, a specially designed Vision Transformer (ViT) built on transformer architecture was also trained. Transformers have demonstrated remarkable performance in a range of computer vision applications, such as picture categorisation. The ViT model uses a vision transformer architecture that captures both local and global dependencies in the picture while processing image patches effectively using self-attention mechanisms. The new design of the model is based on the transformer architecture described in research paper 13. Figure 5 illustrates the leading architecture of this concept, which is as follows:

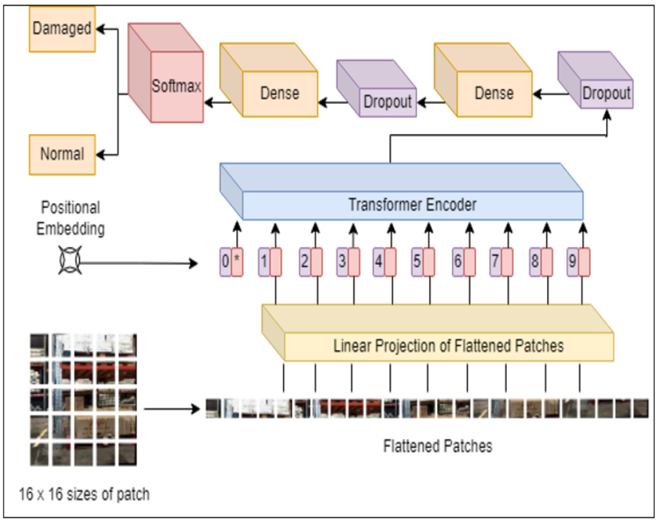

Fig 5 Custom vision transformer architecure.

- *Patch Embedding Layer:*

The image is initially processed with a layer of patch edging. Getting 16x16 patches from the input picture and applying a dense projection to each patch are the main goals at this point. Each patch also includes a class token and positional embedding, which help the model learn about the spatial and semantic elements of the picture.

- *Transformer Encoder Layer:*

After that, the patch edging layer's output is routed through many transformer layers. The model consists of eight Transformer Layer layers, each with a feed-forward neural network and a multi-head self-attention mechanism. The feed-forward neural network assists in obtaining higher-level information from the many picture patches that the model has assigned weights to, thanks to the multi-head attention mechanism. Several dense layers with dropout rates of 0.3 and 0.2 are included after the Transformer encoder layers. In the end, this yields a softmax classifier.

➢ *Compact Convolutional Transformer (CCT)*

A third model, a Custom Compact Convolutional Transformer (CCT), was developed for benchmarking. By using both local and global information, this model enhances feature extraction through the use of compact convolutions and the transformer architecture. Unlike the proprietary ViT model, its main goal is picture classification by feature extraction and transformer-based architecture processing. Based on a study report [37], the Custom Compact Convolutional Transformer model was created. As seen in Figure 6, there are two main construction processes in the CCT model: Transformer and Convolutional Tokenization with Sequence Pooling.

- *Convolutional Tokenizer:*

This section's job is to take features out of the supplied image. A series of convolutional layers with a kernel size of 3, a stride of 1, and a padding of 1 are used to accomplish this. A pooling process is then carried out. A collection of patches, each of which represents a distinct area of the image, is the result of this method. After that, these patches are moved to the Transformer Encoder block to undergo further processing. The function may be expressed using equation (15) [37].

$$available x_0 = MaxPool\left(ReLU\left(Conv2d(x)\right)\right) \qquad (15)$$

Given a feature map or picture where x ∈RH×W×C, where c is the number of channels, w is the weight, and H is the height.

- *Transformer with Sequence Pooling:*

The Transformer Encoder, the initial component of this layer, attempts to comprehend the connections between various patches that the Convolutional Tokenizer block extracts. It contains 64 projection dimensions, eight transformer layers, and four attention heads. A Multi-Head Attention method is used in this section to assist the model in focusing on particular regions of the picture while taking the entire image into consideration. The output is then sent to Sequence Pooling, which pools across the token sequence using an attention-based methodology. This change results in a minor reduction in computation since fewer tokens are being transmitted.

A feed-forward network (MLP) is also used by the Transformer Encoder block to analyse the characteristics that the Multi-Head Attention mechanism has retrieved. The transformer encoder's transformer units are configured to 128. Finally, many FC layers are applied to the Transformer Encoder block's output in order to get the final categorisation. A dense layer with 512 units, another dense layer with 256 units, and a third dense layer with three units make up the FC layers. A SoftMax activation function is used in the last layer to output the probability for each class.

IV. EXPERIMENTAL RESULTS

A. *Experimental Setup*

A laptop with an AMD Ryzen 9 5900HX CPU, 16 GB DDR4 RAM, and an NVIDIA GeForce GTX 3070 with 8GB GDDR6 GPU was used for the research described in this paper. Using Mathplotlib [39], Pandas [40], and Keras [38] from the DL libraries, the Python programmes were created.

B. *Data Partition*

It was essential to separate our dataset into three subsets for testing, validation, and training in order to train our model efficiently. At first, we designated 80% of our photos as part of the training set, while the remaining 20% were kept just for testing. However, in order to guarantee the best accuracy and lower the chance of overfitting, we further divided our training set into two subgroups. Eighty percent of our training photos were used for the actual model training, while twenty percent went into the validation set. With this method, we were able to keep a close eye on our model's performance and modify our training regimen as needed. Following the process, Table 1 displays the number of photos for each region.

Table 1 Different Subsets of Dataset

Data	**Samples**
Training	836
Validation	238
Testing	127

C. Model Hyperparameters and Training Setup

Table 2 displays the hyperparameters used to train each model. Throughout the training phase, 150 epochs of training were permitted for the models. An Early Stopping method was included to avoid overfitting. This function keeps an eye on the validation accuracy of the model and halts training when the accuracy ceases, increasing by at least 1e-4 for a continuous period of 15 epochs. Additionally, while training, the best weights are recovered. This guarantees that the weights of the model from the epoch that performed the best on the validation set will be used for the final assessment.

Table 2 Standard Hyperparameters across all Models.

Hyperparameter Name	**Hyperparameter Value**
Batch Size	32
Learning Rate	0.001
Weight Decay	0.001
Optimizer	Adam

To enhance optimisation and prevent overfitting, the TensorFlow library's CosineDecay function is used as the learning rate scheduler during training [41]. To find the ideal learning rate value, the learning rate is first set at 0.001 and then progressively reduced over little stages. A sharp overshooting and a severe drop in accuracy were noted if the initial learning rate was greater than this number. This method makes sure the model starts out with a high learning rate, which allows it to converge fast, and then progressively lowers the learning rate over time to fine-tune the model. During training, the optimiser's learning rate is updated via the LearningRateScheduler callback.

D. Evaluation

We used the unweighted mean to average the class-wise scores that we calculated in our experimental setup in order to assess the models' performance. We used a number of performance criteria that are widely accepted in the community to evaluate the efficacy of our strategy. The metrics listed below were used.

➢ *Accuracy*

This indicator, which shows the percentage of properly identified cases, assesses how accurate the model's predictions are overall.

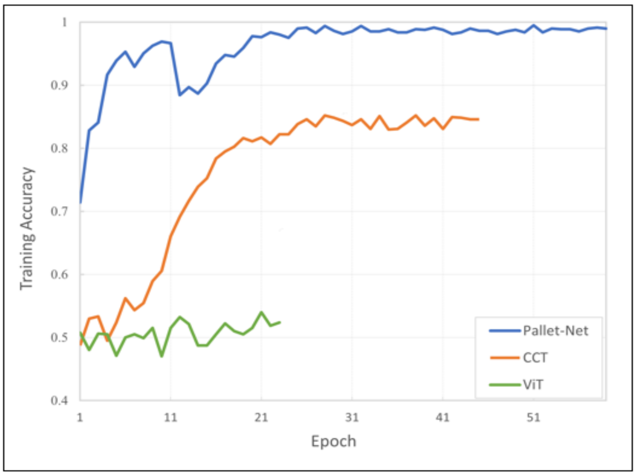

Fig 7 Pallet-Net, vit and CCT Epoch vs training Accuracy

- *This Measure was Computed using Equation (16).*

$$Accurcay = \frac{TP+TN}{TP+FP+TN+FN} \qquad (16)$$

➢ *F1-score*

The model's equilibrium between recall and accuracy is gauged by the F1-score. Equation (17) is used to assess the model's accuracy in classifying both positive and negative events.

$$F1 = \frac{2TP}{2TP+FP+FN} \qquad (17)$$

➢ *Sensitivity (True Positive Rate)*

The percentage of accurate positive predictions among all positive occurrences is known as sensitivity. Equation (18) is utilised to calculate the model's accuracy in identifying positive cases.

$$F1 = \frac{2TP}{2TP+FP+FN} \qquad (18)$$

➢ *False Positive Rate (FPR)*

Out of all negative cases, the fraction of wrongly anticipated positive instances is quantified by the FPR. It gauges the model's propensity to mistakenly identify negative situations as positive. The FPR formula is represented by equation (19).

$$FPR = \frac{FP}{FP+TN} \qquad (19)$$

We employ community-standard performance criteria for Equations (16), (17), (18), and (19) [41] to assess the precision of our models in identifying cases of defective pallet racking. False positive (FP), false negative (FN), true positive (TP), and true negative (TN) are examples of this. By using these indicators, we are able to compare our results with previous methods and gain a thorough understanding of the performance of our model. With this method, we can evaluate our model's performance in a consistent and industry-accepted way.

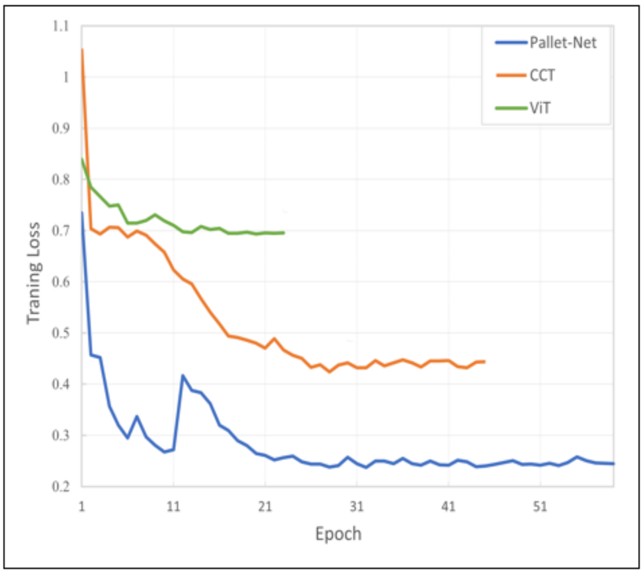

Fig 8 Pallet-Net, vit and CCT Epochs vs training Loss

E. Results

We ran a number of tests to evaluate the effects of the suggested pallet net. Table 3 shows the training duration, parameters, recall, f1-score, precision, and accuracy of our suggested design in comparison to well-known modern architectures such as the Compact Convolutional Transformer (CCT) and Vision Transformer (VIT). Additionally, the models' training accuracy outcomes and training losses are shown in Figures 7 and 8.

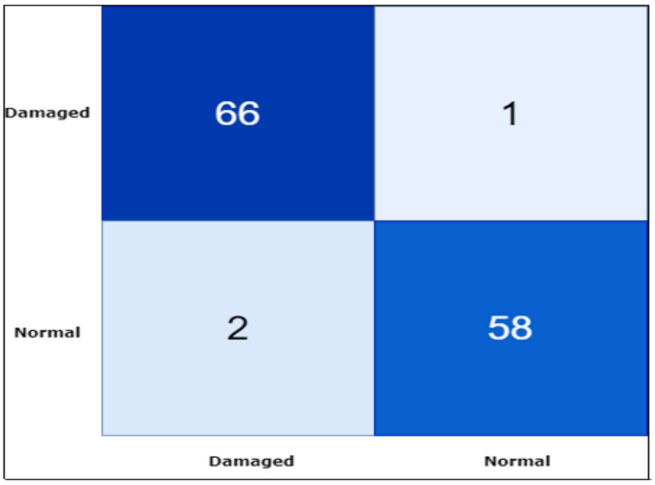

Fig 9 Confusion Matrix of Proposed Model

Table 3 Quantitative Examination and Comparative analysis of Model Performance on test Datase

Model	Training Time	Total Params	F1 Score	Recall	Precision	Accuracy
ViT	03m55s	296066	34%	49%	26%	52%
CCT	05m29s	240451	87%	87%	87%	87%
AttentionCNN	06m24s	154279331	98%	98%	98%	98%

Automated pallet racking classification proves non-trivial, with the Vision Transformer (ViT) architecture demonstrating limited accuracies of around 34% F1 despite low parametrisation and training costs. ViT's inability to aptly capture subtle damage morphology cues limits precision and Recall alike. In contrast, the Compact Convolutional Transformer (CCT) better balances representational complexity and training efficiency, achieving improved 87% F1 classification performance at marginally higher resource overheads. CCT's embedded convolutional feature extraction likely accounts for enhanced localisation of damage signatures within broader rack imagery context to enhance positive and negative instance prediction consistency. Ultimately, our proposed attention-augmented Convolutional Neural Network (CNN) significantly outperforms both baseline approaches, reaching 98% F1-scores, by dedicating a majority of representational capacity towards hierarchical damage characteristics cognition. The additional parameters enable discerning highly complex and variable pallet-racking distortion topologies amid clutter. Our evaluations reaffirm target-specific selectivity as the cornerstone of efficient biological perception and intelligence, which is now gaining traction in biomimetic automated monitoring. Deliberate representation skewing towards explanatory factors, rather than blanket resource scaling, continues to drive innovation.

In order to evaluate how well the suggested Pallet-Net model performed in comparison to the real damage categories, we also created the confusion matrix shown in Figure 9. The classification findings' real positives and negatives, as well as false positives and negatives, are shown in a 2x2 table called the matrix. Correctly categorised data is represented by the diagonal of the confusion matrix, and incorrectly classified data is represented by the off-diagonal components. Pallet-Net properly recognised 66 out of 67 actual damaged racking photos as damaged, according to the confusion matrix. Comparable to the 60 real normal racking photos, just one was incorrectly identified as normal at the same moment. Allet-Net identified two racking photos as damaged but properly identified 58 as normal. With an overall accuracy rating of 97.64%, the suggested model has a high level of accuracy overall.

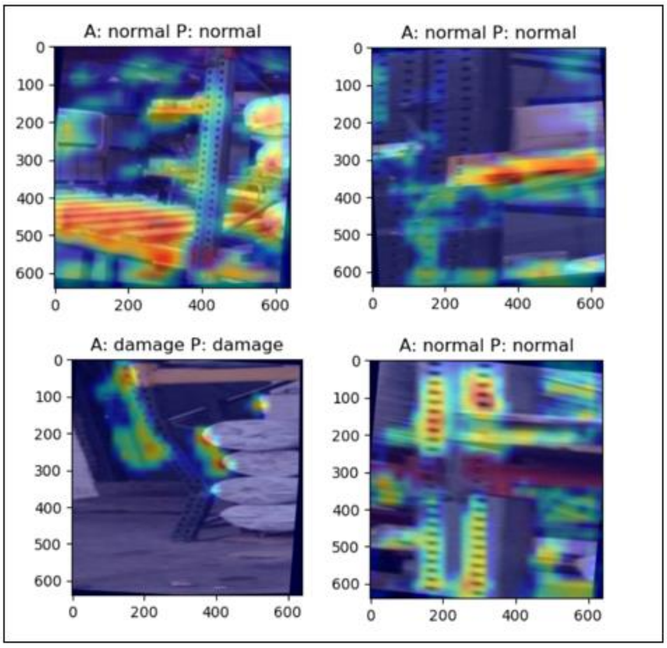

Fig 10 Correctly Classified Cases and their Attention Heatmap via Grad Cam

In Pallet-Net, we have used Gradient-weighted Class Activation Mapping (Grad-CAM) visualisations to identify the important areas of the input photos in order to assess how well the feature extraction method worked. A Grad-CAM depiction of our architecture is shown in Figure 10. Our investigation shows that although Pallet-Net's attention mechanism successfully distinguishes between damaged and undamaged racking by identifying the critical structures of the racking, it occasionally focuses on the pallet region and other non-salient image regions, which could lead to incorrect classification. As a potential remedy, we advise applying preprocessing methods to improve Pallet-Net's accuracy, such as filtering out unnecessary or irrelevant areas.

It is clear from the technical evaluation that the pallet net performs better in terms of classification accuracy than the ViT and CCT models. Greater performance is achieved by the models with more parameters and longer training sessions because they show a deeper comprehension of the intricate linkages present in the damaged pallet racking photos.

V. SOLUTION COMPARISION

Automated analytical workflows aim to balance accuracy, efficiency and accessibility for real-world damage monitoring integration. Farahnakian et al.'s [42] segmentation methodology demonstrates leading 93.45% precision but on highly constrained datasets given intensive resource demands. Conversely, Hussain et al.'s [21] YOLOv7 detector attains 91.1% accuracy on thousands of samples, although at the cost of additional bounding box annotations and computations. Recent focus has shifted to streamlined classification via lightweight architectures [43], reaching 96% accuracy under reasonable resource profiles. However, despite higher expenses, MobileNet-powered detection [19] remains competitive, indicating scenarios where detail surmounts efficiency.

Our attention-based classifier achieves an accuracy of 97.63% on over a thousand pallet rack images lacking supplementary bounding boxes, setting new state-of-the-art performance. Compared to prevailing techniques, our methodology promises a pragmatic balance of damage cognisance, computational frugality and real-world validity for scalable rack monitoring autonomy. More broadly, the comparative analysis spotlights representational selectivity as the lingering bottleneck for pervasive intelligence. While sheer analytical muscle continues steadily improving, deliberate dimensionality pruning to amplify explanatory factors over superfluous imagery traits remains crucial but underexplored. As domains such as biomarker discovery already underscore, sparsity frequently surpasses scale for unravelling complex phenomena. Our evaluations reaffirm this motif - superior cognition arises from compact, causal models rather than indiscriminate resource intensification.

Table 4 Systematic Evaluation and Comparative Analysis with Prior Research in the field

Research	Domain	Dataset Size	Detector	Accuracy
[43]	Image Classification	1723	Custom CNN	96%
[42]	Segmentation	75	Mask RCNN	93.45%
[19]	Object Detection	19717	Mobile Net	92.7%
[21]	Object Detection	2094	YOLOv7	91.1%
Proposed	Image Classification	1201	Attention CNN	97.63%

In summary, compared to other studies on automated racking inspection, Pallet-Net, the suggested attention-based CNN architecture, offers better accuracy and a simpler processing pipeline. It provides a more dependable and effective way to identify and categorise damage to pallet racking, allowing the warehouse sector to operate with more efficiency, lower costs, and higher safety.

VI. CONCLUSION

This research pioneers Pallet-Net - an attention-focused convolutional neural network (CNN) architecture achieving automated state-of-the-art pallet racking damage detection at 97.64% accuracy. We systematically enhance representation learning using grayscale conversion, image resizing and data augmentation that exposes models to real-world environmental complexity while steeping them specifically in damage morphology. Our tailored CNN then develops hierarchical damage characterisations amplified by integrated attention mechanisms highlighting spatial irregularities. Comprehensive evaluations versus contemporary Vision Transformer and Compact Convolutional Transformer architectures reaffirm attention's efficacy for potent yet selective rack cognition. Pallet-Net promises efficient automation unattained by blanket computational scaling or human visual assessment alone. More broadly, it epitomizes an awareness amplification motif gaining traction across biomedicine, manufacturing, and more - seemingly boundless societal challenges are increasingly yielding not to brute analytical force but deliberate, causal representations distilling phenomena down to their essence. As datasets now expand worldwide, scalable intelligence will arise from carefully tuned filters revealing what truly matters.

This research pioneers automated pallet racking assessment via selective deep learning, surpassing constrained human visual scrutiny. Pallet-Net exemplifies augmented cognition - not brute analytical force alone - achieving previously unattained warehouse visibility. Our framework promises enhanced safety, efficiency and risk attenuation beyond current practice. We acknowledge sample size limitations among other constrained resources typical of initial investigations now outpacing isolated human perspective. Ongoing efforts will enrich representations and explore modern architectures. Ultimately, damage detection applies the selectivity gaining prominence from healthcare to renewables. Embedded intelligence that amplifies the most explanatory cues in environments otherwise overwhelming human operators must emerge. As automation broadly displaces specialised operators and sensors, next-generation methods embedding extracted wisdom into key processes promise democratised situation awareness, benefiting society widely. Scalable and reliable intelligence resides in deliberate representations - the essence revealed matters more than the resources invested. Our research manifests this new paradigm centred on awareness rather than just analysis.

REFERENCES

[1]. Bernuzzi, M. Simoncelli, An advanced design procedure for the safe use of steel storage pallet racks in seismic zones, Thin-Walled Structures 109 (2016) 73–87.

[2]. M.Hussain, H. Al-Aqrabi, M. Munawar, R. Hill, S. Parkinson, Exudate regeneration for automated exudate detection in retinal fundus images, IEEE access(2022)1doi:https://doi.org/10.1109/access.2022.3205738.

[3]. B. A. Aydin, M. Hussain, R. Hill, H. Al-Aqrabi, Domain modelling for a lightweight convolutional network focused on automated exudate detection in retinal fundus images, in: 2023 9th International Conference on Information Technology Trends (ITT), IEEE, 2023, pp. 145–150.

[4]. Zahid, M. Hussain, R. Hill, H. Al-Aqrabi, Lightweight convolutional network for automated photovoltaic defect detection, in: 2023 9th International Conference on Information Technology Trends (ITT), IEEE, 2023, pp. 133–138.

[5]. M. Hussain, H. Al-Aqrabi, M. Munawar, R. Hill, Feature mapping for rice leaf defect detection based on a custom convolutional architecture, Foods 11 (23) (2022) 3914. doi:10.3390/foods11233914.

[6]. K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[7]. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Report, California Univ San Diego La Jolla Inst for Cognitive Science (1985).

[8]. K. Simonyan, A. Zisserman, Very deep convolutional networks for largescale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[9]. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10]. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1–9.

[11]. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 580–587.

[12]. R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[13]. C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, Yolov7: Trainable bagof-freebies sets new state-of-the-art for real-time object detectors, arXiv preprint arXiv:2207.02696 (2022).

[14]. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE signal processing magazine 35 (1) (2018) 53–65.

[15]. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.

[16]. X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, Pre-trained models: Past, present and future, AI Open 2 (2021) 225–250.

[17]. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[18]. H.-H. Zhu, F. Dai, Z. Zhu, T. Guo, X.-W. Ye, Smart sensing technologies and their applications in civil infrastructures 2016 (2016).

[19]. M. Hussain, T. Chen, R. Hill, Moving toward smart manufacturing with an autonomous pallet racking inspection system based on mobilenetv2, Journal of Manufacturing and Materials Processing 6 (4) (2022) 75.

[20]. C.-Z. Dong, F. N. Catbas, A review of computer vision–based structural health monitoring at local and global levels, Structural Health Monitoring 20 (2) (2021) 692–743.

[21]. M. Hussain, H. Al-Aqrabi, M. Munawar, R. Hill, T. Alsboui, Domain feature mapping with yolov7 for automated edge-based pallet racking inspections, Sensors 22 (18) (2022) 6927.

[22]. S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, B. Lee, A survey of modern deep learning based object detection models, Digital Signal Processing (2022) 103514.

[23]. Y. Liu, P. Sun, N. Wergeles, Y. Shang, A survey and performance evaluation of deep learning methods for small object detection, Expert Systems with Applications 172 (2021) 114602.

[24]. J. Yang, S. Li, Z. Wang, G. Yang, Real-time tiny part defect detection system in manufacturing using deep learning, IEEE Access 7 (2019) 89278–89291.

[25]. T. D. Akinosho, L. O. Oyedele, M. Bilal, A. O. Ajayi, M. D. Delgado, O. O. Akinade, A. A. Ahmed, Deep learning in the construction industry: A review of present status and future innovations, Journal of Building Engineering 32 (2020) 101827.

[26]. D. Weimer, B. Scholz-Reiter, M. Shpitalni, Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection, CIRP annals 65 (1) (2016) 417–420.

[27]. H. Liu, L. Zhou, J. Zhao, F. Wang, J. Yang, K. Liang, Z. Li, Deeplearning-based accurate identification of warehouse goods for robot picking operations, Sustainability 14 (13) (2022) 7781.

[28]. Z. Tang, E. Tian, Y. Wang, L. Wang, T. Yang, Nondestructive defect detection in castings by using spatial attention bilinear convolutional neural network, IEEE Transactions on Industrial Informatics 17 (1) (2020) 82–89.

[29]. B. Su, H. Chen, P. Chen, G. Bian, K. Liu, W. Liu, Deep learning-based solar-cell manufacturing defect detection with complementary attention network, IEEE Transactions on Industrial Informatics 17 (6) (2020) 4084–4095.

[30]. T. B. Shahi, C. Sitaula, A. Neupane, W. Guo, Fruit classification using attention-based mobilenetv2 for industrial applications, Plos one 17 (2) (2022) e0264586.

[31]. Chollet, Building powerful image classification models using very little data, Keras Blog 5 (2016) 90–95.

[32]. D. Singh, B. Singh, Feature wise normalization: An effective way of normalizing data, Pattern Recognition 122 (2022) 108307.

[33]. Al-Sadi, A.-A. M. Hana'Al-Theiabat, M. Al-Ayyoub, The inception team at vqa-med 2020: Pretrained vgg with data augmentation for medical vqa and vqg, in: CLEF (Working Notes), 2020.

[34]. Z. Hussain, F. Gimenez, D. Yi, D. Rubin, Differential data augmentation techniques for medical imaging classification tasks, in: AMIA annual symposium proceedings, Vol. 2017, American Medical Informatics Association, 2017, p. 979.

[35]. L.-F. Li, X. Wang, W.-J. Hu, N. N. Xiong, Y.-X. Du, B.-S. Li, Deep learning in skin disease image recognition: A review, IEEE Access 8 (2020) 208264–208280.

[36]. M. A. R. Alif, S. Ahmed, M. A. Hasan, Isolated bangla handwritten character recognition with convolutional neural network, in: 2017 20th International conference of computer and information technology (ICCIT), IEEE, 2017, pp. 1–6.

[37]. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, H. Shi, Escaping the big data paradigm with compact transformers, arXiv preprint arXiv:2104.05704 (2021).

[38]. N. Ketkar, N. Ketkar, Introduction to keras, Deep learning with python: a hands-on introduction (2017) 97–111.

[39]. Bisong, E. Bisong, Matplotlib and seaborn, Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners (2019) 151–165.

[40]. W. McKinney, pandas: a foundational python library for data analysis and statistics, Python for high performance and scientific computing 14 (9) (2011) 1–9.

[41]. D. So, Q. Le, C. Liang, The evolved transformer, in: International conference on machine learning, PMLR, 2019, pp. 5877–5886.

[42]. Farahnakian, L. Koivunen, T. Makil¨ a,¨ J. Heikkonen, Towards autonomous industrial warehouse inspection, in: 2021 26th International Conference on Automation and Computing (ICAC), IEEE, 2021, pp. 1–6.

[43]. M. Hussain, R. Hill, Custom lightweight convolutional neural network architecture for automated detection of damaged pallet racking in warehousing & distribution centers, IEEE Access (2023).