Application Research and Analysis of Computer Vision in the Field of Construction Engineering

Junzhao Huang[1*]

[1]Department of Computing, Harbin Institute of Technology,150001, Harbin, Heilongjiang Province, China

Abstract. Aiming at the problems of low degree of automation, frequent safety accidents and difficult quality control in construction engineering, this paper studies the application of computer vision, and focuses on the core idea of algorithm, underlying logic, current application and future development of Yolo series, SC crack+and efficientnet models. Yolo is good at real-time multi-target detection, which is suitable for component management and crack detection; SC crack+realizes accurate crack segmentation in complex environment through unsupervised mechanism, making it an ideal choice for traditional buildings with scarce samples. Efficientnet stands out in image classification and multimodal data fusion through composite scaling and lightweight design. Challenges remain, including Yolo's limited small target accuracy, the hyperspectral data cost of SC crack+and the generalization limitation of efficientnet. The improvement directions such as multi-scale feature fusion, spatial constraint optimization and multi-modal data integration are proposed to support the intelligent transformation in construction monitoring and structural health assessment.

1 Introduction

Construction industry is an important part of the global economy, and it has traditionally experienced the lack of automation, frequent accidents concerning safety, and problems in quality control. It is still comparatively slow in its digital and intelligent transformation because of the intricate nature of construction projects, their scale and dynamism. Manual inspection and documentation that is still being used in traditional construction management is inefficient, subjective and ends up producing incomplete data and low accuracy. Such constraints often cause cost increases, time wastages and quality and safety risks that are not detected. Thus, the invention of higher technical capabilities to build more intelligence in the company engineering is an acute industrial need [1,2].

Computer vision is a crucial field of the artificial intelligence that replicates the computer visual system to gather, process and analyze image and video data. It is a natural and viable technology route upon which the construction industry can be converted and transformed and modernized. As of recent years, computer vision has proven to perform excellently with great

* Corresponding author's email: 2023113357@stu.hit.edu.cn

progress in the field of image recognition, object detection, and instance segmentation with notable progress in the deep learning algorithms [3].

Already three computer vision models have been used with proper and effective applications in construction engineering. Yolo is capable of real time detection, which is excellent; thus, it is well positioned in building surface defect recognition and prefabricated component control [4]. SC Crack+ model is specially created to detect cracks and monitor the structural health and has high flexibilities to the images taken in complex conditions [5]. EfficientNet which is effective and efficient has widely been utilized in construction monitoring, building identification and infrastructure management [6,7].

The given paper is aimed at the comparison of the above three models, their area of application, practical utility, and possible optimization solutions.

2 Key Technologies and Models

2.1 Yolo Series Models

2.1.1 Core Idea of Yolo Algorithm

The Yolo framework reframes object detection as a single regression problem, enabling end-to-end prediction of bounding boxes and class probabilities directly from input images in a single forward pass [4]. As a one-stage detector, it eliminates the need for region proposal generation, resulting in a simple and fast pipeline. Its loss function is designed to minimize regression error within a convolutional neural network, directly optimizing for localization and classification. This end-to-end architecture delivers exceptional real-time performance while maintaining competitive accuracy. Through continuous improvements in network design and training strategies, Yolo has achieved an effective balance between speed and precision, making it widely adopted in real-time detection tasks, including small-object identification.

2.1.2 Iterative Improvement of Yolo Series Models

Proposed by Redmon et al. the choice of the original Yolo model transformed object detection into a single regression problem that not only fundamentally altered the process of object detection but also significantly accelerated it [4]. Yolo V2 and V3 now have anchor boxes, multi-scale prediction, and less efficient backbone networks (Darknet-19/53) respectively, and have much improved their performance. The groundwork of the Yolo series provided a very strong basis upon which breakthroughs were to be made in the future.

Yolo V5 had also mosaic data augmentation, a new anchor-free detection head, and is written in PyTorch, quickly becoming a standard in the industry used in practice. Yolo V7 was the first one to make significantly radical architectural changes that made considerable improvements to accuracy and efficiency. Yolo V8 represents an integrated model, which allows detection, segmentation, pose estimation, and other visual AI [4]. It embraces a novel backbone network (C2f organization), decoupled, anchor-free head, and task alignment assigner.

Yolo has now been iterated to the Yolo V11. Not only is the latest version guaranteeing real-time performance, but also enhancing the accuracy of detection and the ability of feature extraction with reasonable architectural optimisation introducing C3K2, SPPF and C2PSA modules. Hence, it has intuitively become a highly mature and dependable option of industrial detection [4].

2.2 Sc-Crack+ Crack Segmentation Model

SC Crack+ is an automatic way of dividing crack in a hyperspectral image [5]. It then uses the Spearman rank correlation coefficient to automatically remove spectral channels causing redundancy in the bands, but preserves discriminative features. The pixels are then clustered into small clusters using K- means clustering and the Euclidean distance between the center of the clusters and the spectral origin is calculated. These distances are normalized, depending on the low reflectance characteristics of regions in cracks and a fixed threshold is used to generate a crack super-cluster, which allows constructing a stable segmentation without tuning the parameter by hand.

SC Crack+ has specific advantages over the traditional threshold-based or deep learning methods. It does not depend on an annotated training data and is entirely unsupervised, which makes it a particularly appropriate choice when it comes to heritage buildings and other contexts where annotated training data of large scale cannot be obtained. Besides, the far-infrared spectral information will be used to distinguish not only the cracks but also the biological ones like vegetation, moss and lichen, which can be divided using the near-infrared spectral information and will decrease the false alarms substantially [6]. This automatic channel selection mechanism does not only boost the segmentation performance but then also instructs the simplified design of the future hyperspectral imaging systems.

2.3 EfficientNet Series Models

EfficientNet is a convolutional neural network that is an architecture concerning performance and computational efficiency such that it considers both the performance and the computational efficiency so it has very clear and excellent characteristics [7]. First is the compound scaling principle in which the network depth, width and input image resolution are reasonably and well balanced adjusted and can produce best performance in situations where there are limited computing resources. Second, its simplest module MBConv (Mobile Inverted Residual Bottleneck Convolution) is a combination of the depthwise separable convolution and Squeeze-and-Excitation attention mechanism in which the network draws attention to the important elements in the image of the architectural objects entirely and correctly, and therefore it is highly appropriate at extracting the features of the complex texture, fine structure and numerous variations in the image of the architectural scenes. Third, it is very good in terms of transfer learning. In the applications of the model, at the level of solving of the specific tasks in the construction engineering (finding cracks in the image, identification of the elements of the building, etc.), when trained using the ImageNet large image dataset, can nevertheless successfully and quickly migrate to the case of extremely limited data on professional annotations [6] then it turns out to be a highly potent and convenient visual feature extractor to the visual problems of construction engineering.

3 Application

3.1 Integration and Promotion Application of Yolo V11

Aiming at the problems of feature information loss, low detection accuracy of small targets, and false detection and missed detection in complex scenes, researchers optimized Yolov11 from three aspects: backbone network, feature fusion module, and loss function [4]. First, the C3K2-SG module, which integrates SMPConv and CGLU, is introduced into the backbone network to suppress redundant responses through the point movement mechanism and enhance the expression ability of local crack characteristics in complex backgrounds. Second, the FPSConv module is designed to replace SPPF. Shared convolution with different expansion factors (1, 3, 5) is used to extract multi-scale features, and channel fusion is realized through 1×1 convolution, which improves the recognition effect of micro-cracks while reducing the number of parameters. Finally, the Inner-MPDIoU loss function is proposed. Based on MPDIoU, a scale factor is introduced and combined with the Inner-IoU adaptive mechanism to improve the localization accuracy and model convergence speed for small-target cracks.

Experimental results show that the improved Yolov11 achieves an mAP@0.5 of 88.6% in building crack detection tasks, which is 4.6% higher than the original model, while maintaining real-time performance at 212 FPS. It demonstrates significant advantages in detecting small targets and complex crack scenes, providing an efficient and reliable detection scheme for building structure safety monitoring [4].

3.2 Application of Sc Crack+ Crack Segmentation Model

The model finds numerous applications in real engineering based on the unsupervised trait of SC Crack+ and strong capability of operating in complex backgrounds.

First, it can be co-opted in the regular inspection of tunnels, retaining walls, and concrete walls, and urban infrastructures by using vehicle-mounted or UAV hyperspectral imaging systems to achieve automatic identification and location of massive cracks in the concrete [5]. Secondly, the model could be an front-end sensing device in smart structural health monitoring (SHM) systems, which would allow the provision of a high-quality input to continue crack width evaluation, disease progression, and life estimation.

Moreover, the marine buildings or structures that may be located in high-humidity conditions, the ability of the SC Crack+ to protect against biological attachments becomes an advantage of application in the situations when the use of the traditional visible light is not efficient. Lastly, as the method can run without the use of training data, it can be used in resource-constrained or real-time detection systems (where it has high engineering practicability) and it can be promoted (where there has been strong engineering practicability).

3.3 Application of EfficientNet Series Model

The scope of application of EfficientNet within the area of the construction engineering is bound to be extended further due to the ongoing growth of technology. EfficientNet should be self-integrated with graph neural network (GNN) and other models in the future. It is able to utilize EfficientNet to derive visual features, Transformers to compute spatial relationships between parts of the building, and GNNs to compute topological structures of a building information model (BIM). Such a combined methodology makes multi-modal data processing collaborative [8,9,10].

The framework is capable of integrating directly the image data with multimodal data including laser point clouds, thermo imaging, and BIM data, where EfficientNet is in charge of the visual part. It will be possible to analyze the construction project progress, quality, and energy consumption comprehensively and contribute to more adequate and more accurate digital twin modeling and smart operation and maintenance.

Moreover, edge computing in real time is also a good fit with the model. It is extremely beneficial in terms of placing it on edge devices (i.e., UAVs, inspection robots, cameras in construction sites, etc.) because of its lightweight nature, which allows the system to respond within milliseconds to real-time monitoring and provide early warning [6].

4 Conclusion

Dedicating the smart needs of the construction engineering, this paper conducts a systematic study on the utilization of computer vision technology in construction engineering, and put particular attention to the technical peculiarities, engineering possibilities, and the optimization prospects of the Yolo series, SC Crack+, and EfficientNet models.

According to the research, Yolo series is sensitive to real time and multi-target detection and thus can be used in controlling prefabricated components and crack finding. In complex environments, SC Crack+ can generate accurate distribution of the cracks using the unsupervised mechanism which is particularly applicable in situations where there are limited samples. EfficientNet stands out by classifying images and combining different modalities, as well as scaling compounds in a design of compound scaling.

Nevertheless, current models also have a number of limitations, among which is the lack of accuracy of Yolo required to detect small targets, a high cost of hyperspectral data of SC Crack+, and a weak generalization performance of EfficientNet. The improvement of the model robustness and practical application in complex engineering applications by the incorporation should be done in the future, using strategies like multi-scale feature fusion, the optimization of the spatial constraints, and multimodal data integration, so that it will offer technical support to the intelligent transformation of construction engineering.

References

1. W. Lu, et al. Computer vision-based construction and demolition waste identification. Autom. Constr. (2025)
2. A. Abbasi, et al. Roof material detection using aerial imagery and deep learning. Sustain. Cities Soc. (2024)
3. M. Toğaçar, et al. Deep learning-based waste classification using autoencoder and CNN. Appl. Sci. (2024)
4. R. Magdy, K.A. Hamdy, Y.A.S. Essawy. Real-time progress monitoring of bricklaying using computer vision. Buildings, 15, 2456 (2025)
5. B.O. Santos, J. Valença. Relevant channel selection in hyperspectral imaging to enhance crack segmentation in historic concrete buildings. Infrared Phys. Technol., 150, 105958 (2025)
6. D.Rajaratnam, R.A. Stewart, T. Liu, A.S. Vieira, M.E.U. Zannat. Computer vision-based approach for autonomous estimation of building envelope material stock. Sustain. Cities Soc., 136, 107049 (2026)
7. M. Farshadfar, et al. AI and machine learning for construction waste management: a review. Sustainability (2025)
8. R. Moya-Jiménez, A. Goyes-Balladares, G. Moya-Jiménez, et al. Application of computer vision and parametric design algorithms for the reuse of construction materials. Buildings, 16, 184 (2026)
9. X. Liu, et al. Artificial intelligence-based construction waste recognition and sorting. J. Clean. Prod. (2025)
10. Y. Dong, et al. Boundary-aware transformer for construction waste segmentation. Autom. Constr. (2025)