International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 284-294

https://doi.org/10.70517/ijhsa46423

Application of Smart Media Fusion Technology Based on Image Recognition and Computer Vision Algorithms in the Digital Conservation and Display of the “Gongshu Hall”

Qi Yin[1] and Hongfei Chang[2,3,*]
[1] College of Art and Design, Xi'an Innovation College of Yan'an University, Xi’an, Shaanxi, 710100, China
[2] College of Humanities and Social Sciences, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China
[3] College of Design and Art, Xijing University, Xi'an, Shaanxi, 710123, China
Corresponding authors: (e-mail: 15353626773@163.com).

Abstract Digital protection and display of ancient buildings is an important technical means of cultural heritage transmission, and “Gongshu Hall” as a typical wooden ancient building, its wall materials are complex and diverse, and the precise identification and protection of surface damage is in urgent need. This study proposes a set of intelligent media fusion technology framework based on image recognition and computer vision algorithms, and realizes high-precision material recognition and damage detection by optimizing the network structure, loss function and algorithm fusion strategy. Firstly, the EfficientNet v2 network is improved, and the collaborative attention mechanism (CA) is introduced to replace the original SE module to enhance the spatial location perception of the feature map. To solve the problem of insufficient bounding box regression accuracy of YOLOv7 in crack detection, the sample gradient contribution is balanced by the normalization factor and monotonic focusing coefficient, which improves the model convergence speed and location accuracy. The two-level detection-segmentation joint algorithm is further constructed by combining the pixel-level segmentation capability of UNet3+ network. The model achieves a test accuracy of 93.87% on a dataset containing eight types of materials, with the highest recognition rate of metal (97.04%), followed by blue brick (95.13%) and stone (95.27%), but rammed earth (89.72%) and glazed glass (89.43%) are misclassified due to the complexity of surface features. Experiments show that the algorithm has excellent comprehensive performance in the detection of "spalling", "phthalate" and "crack", with an average F1 value of 97.21%, of which the F1 value of crack detection is the highest (97.64%), and the spalling accuracy (99.47%) and phthalate recall rate (95.96%) are outstanding.

Index Terms image recognition, computer vision, EfficientNet v2 network, YOLOv7, UNet3+, “Gongshu Hall”, ancient building damage recognition

I. Introduction

With the rapid development of technology, digitization has become an important means of cultural heritage preservation, especially in the field of architecture. Facing the dangers of deterioration of building materials, gradual loss of construction techniques, and the gradual disappearance of buildings in the process of modernization, many countries are exploring how to preserve and pass on their cultural heritage through digitization technology [1]-[3]. Through digitization, historical buildings are no longer just physically present, but can also continue to be brought to life in a virtual environment.

3D scanning technology can accurately capture the shape, texture and details of a building [4]. Converting buildings into 3D models facilitates observation and analysis anytime, anywhere, archives historical information, and provides important data for subsequent maintenance and restoration, and the virtual display of 3D models enables visitors to have a more intuitive understanding of these cultural heritages, stimulating the public's sense of participation and conservation awareness [5]-[7]. In addition to 3D scanning, the rapid development of Virtual Reality (VR) and Augmented Reality (AR) technology also provides new perspectives for the conservation of architectural heritage. Through VR technology, users can fully immerse themselves in the virtual world of historical buildings, experience different historical periods, and observe the architectural constructions and styles at that time [8]. AR technology, on the other hand, can superimpose virtual information into the real world through intelligent devices, so that users can obtain richer historical background and restoration data even during field visits, and this immersive experience not only provides educational possibilities, but also makes people cherish and respect their cultural heritage more [9], [10].

According to the official data released by the city of Xi'an, Shaanxi Province, the “Gongshu Hall”, built in the Ming Dynasty, suffered serious damage during the Cultural Revolution, and so far the small wood carvings and translucent carvings have suffered significant insect damage, and the coloring facilities faded quickly, and the architectural craftsmanship is exquisite, and the details of the mortise and tenon joints of the small woodwork are worth recording. As the only remaining small woodwork ancient building in China, it is urgent to digitize and protect the“Gongshu Hall”[11]. However, 3D and basic mapping techniques are unable to depict the small details of small woodwork mortise and tenon structures, and the degree of architectural detail restoration is far from 90% [12], [13].

This study combines image recognition and computer vision algorithms to propose a set of intelligent technology framework for the protection of“Gongshu Hall”, and realizes high-precision material recognition and damage detection by optimizing the network structure, loss function and algorithm fusion strategy. Firstly, based on the improved EfficientNet v2 network, a wall material recognition algorithm incorporating the cooperative attention mechanism (CA) is proposed. Aiming at the problem of complex texture and strong spatial information dependence of ancient building materials, the spatial location-awareness ability of the feature map is enhanced by replacing the SE module in the original network with the CA module, and the Adam optimizer is combined to accelerate the convergence of the model, which significantly improves the material classification accuracy. Then for the problem of insufficient bounding box regression accuracy of YOLOv7 algorithm in crack detection, a WIOUv2 loss function with dynamic focusing mechanism is designed. By introducing the normalization factor and monotonic focusing coefficient to balance the gradient contribution of high and low quality samples, the convergence speed and localization accuracy of the detection model are effectively improved. And the UNet3+ network is used to realize the pixel-level segmentation of cracks, whose dense block-and-jump connection structure makes full use of the multi-scale features and solves the mis-segmentation problem caused by background noise in traditional segmentation methods. Finally, YOLOv7 and UNet3+ are further integrated to propose a two-level detection-segmentation joint algorithm.

II. Optimization and integration of intelligent material identification and damage detection algorithms for digital protection

II. A. Improved wall material recognition algorithm based on Efficient Net v2

II. A. 1) Efficient Net v2 network

The new lightweight convolutional neural network Efficient Net v2 differs from traditional convolutional neural networks in that Efficient Net v2 simultaneously scales the three dimensions of the network, namely, width w, depth d, and input image resolution r, instead of adjusting the input image resolution, network depth, or number of convolutional channels individually to improve the network accuracy. This composite scaling method allows the scaled network to improve the model recognition accuracy while minimizing the memory overhead.

Efficient Net v2 has the following improvements over Efficient Net v1:

(1) Efficient Net v2 uses Fused-MBCConv module in addition to MBCConv module at the shallow layer of the network to reduce the number of model parameters.

(2) Remove the last stage with step 1 in Efficient Net v1 due to its excessive number of parameters and memory access overhead.

(3) Efficient Net v2 uses smaller expansion ratios (e.g., the first expansion conv1×1 in MBCConv or the first expansion conv3×3 in Fused-MBCConv is 4, whereas it is basically 6 in Efficient Net v1), which can effectively reduce the memory access overhead.

(4) Efficient Net v2 prefers to use a smaller kernel size (3×3), replacing the more 5×5 kernel size in Efficient Net v1. Due to the smaller sense field, more layers of structures need to be stacked to increase the sense field.

II. A. 2) Algorithm improvement program

The MBCConv modules use depth-separable convolution instead of standard convolution, which constitutes the 4-6 Stage in Efficient Net v2. Each MBCConv module contains the SE module, which is used to efficiently solve the problem of information loss caused by the equal importance of different channels in the feature map. However, the method only considers the information encoding between channels and ignores the important role of spatial information in the recognition of ancient building wall materials, which affects the recognition performance of the model.

The collaborative attention mechanism (CA) module is an efficient attention module that can be easily and flexibly embedded into classical classification networks without adding a large computational overhead, thus enhancing the feature representation capability of convolutional neural networks. This is different from earlier attention mechanisms such as Squeeze-Excitement (SE) that only apply different weights to different channels while ignoring the positional feature information, the CA module uses an efficient method to obtain channel relationship and positional information to achieve better feature representation. Specifically, the CA module aggregates input features in the vertical and horizontal directions into two perceptual feature mappings of independent directions by decomposing the 2D globally pooled feature encoding in the SE module into two parallel 1D feature encodings. This approach generates two attention maps embedded with directional information, which not only captures long-term dependencies along one spatial direction, but also retains precise positional information along the other spatial direction.The specific operation process of the CA module is shown in Fig. 1.

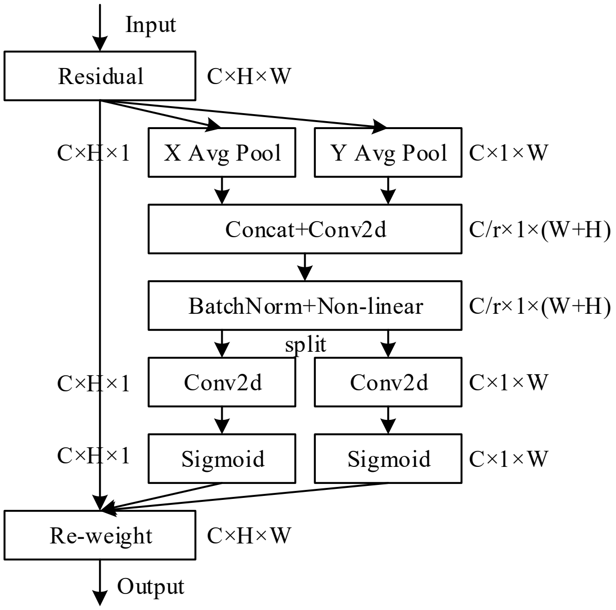


Figure 1: Coordinate A ttention Mechanism Module

Let the input feature map be $F \in R^{C \times H \times W}$, where R denotes the set of real numbers, C denotes the number of channels, H denotes the height of the feature map, and W denotes the width of the feature map.The CA module first decomposes the two-dimensional global pooling operation and transforms it into two one-dimensional global pooling feature encoding operations. Next, the spatial dimensions of the feature information in the two directions are spliced and the feature information is integrated using a 1×1 convolutional transform function to obtain $F' \in R^{(C/r) \times (W+H)}$, which represents the intermediate feature map that encodes spatial positional information in both the horizontal and the vertical directions; where r is a scaling factor used to reduce the number of channels. Subsequently, F' is decomposed into two separate feature maps along the spatial dimensions of the two directions, and the final feature maps G^h and G^w are obtained by two 1×1 convolutional transform functions followed by a Sigmoid activation function. Finally, the input feature map F and the feature maps G^h and G^w in the two directions where the position information is acquired are multiplied by an element-wise multiplication operation to obtain the enhanced feature expression of the CA module as shown in Equation (1).

$$Y' = F \cdot G^h \cdot G^w \tag{1}$$

In order to further improve the accuracy of the recognition model, this paper uses EfficientNet v2-m as the base network and introduces Coordinate Attention module for improvement. In order to avoid the number of model parameters becoming too large and the network inference speed becoming low, the number of module placements cannot be too large. At the same time, due to the shallow module structure is too simple, the extracted features are not enough, the network can not reason well about the position and channel information that need attention. The placement of the collaborative attention mechanism module must not be too shallow. So this paper replaces the SE module in the MBCConv module in Stage6 in the Efficient Net v2 structure with the Coordinate Attention module. Research on mobile network design has demonstrated the significant effectiveness of channel attention (e.g., squeeze-excite attention mechanisms) in improving model performance, and although attention mechanisms are very useful in generating spatially-selective attention maps, they still typically ignore positional information. In this paper, we will insert a collaborative attention mechanism module into the model to embed location information into channel attention. In contrast to channel attention, which only transforms the feature tensor into individual feature vectors through 2D global pooling, coordinate attention decomposes channel attention into two 1D feature encoding processes that aggregate features along two spatial directions, so that long-range dependencies and precise location information can be captured at the same time. The feature maps obtained next are encoded into a pair of direction-aware and location-sensitive attention maps, which can enhance the representation of the object of interest. The added attention mechanism is schematically shown in Fig. 2.

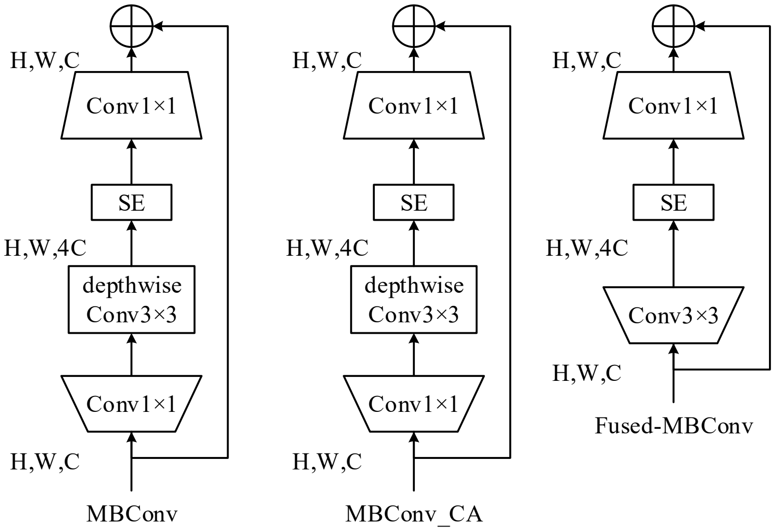


Figure 2: Schematic diagram of adding CA attention mechanism

Efficient Net v2 initially uses the traditional stochastic gradient descent (SGD) method, but in this method, each weight uses the same learning rate, so it is difficult to choose an appropriate learning rate, and the SGD algorithm is easy to fall into the local optimal solution, which leads to a decrease in the recognition accuracy of the model. In order to overcome these problems and speed up the convergence of the model, this paper will use the Adam optimization algorithm in Efficient Net v2. Adam is an adaptive learning optimization algorithm, which not only possesses the advantages of AdaGrad that is good at dealing with sparse gradients, but also combines the advantages of RMSProp that is very good at dealing with non-smooth objectives.

II. B.Boundary box regression loss function based on dynamic focusing mechanism

The accurate identification of wall materials provides basic data support for subsequent damage detection. However, relying only on material classification cannot meet the localization needs of cracks and other damages in ancient building protection, and the bounding box regression accuracy of the target detection algorithm needs to be further optimized.

The loss function of the bounding box regression of the YOLOv7 algorithm is mainly used to measure the degree of difference between the position of the target prediction box and the position of the real box. In training, the closer the predicted frame is to the real frame, the smaller the loss function value is, and the more accurate the prediction result is.

The bounding box regression loss function of the YOLOv7 algorithm uses CIOU by default, which takes into account the influence of many factors on the regression loss value, but does not consider the linear proportion of the height-to-width ratio of the predicted box and the real box, the value of the penalty term becomes 0, which will make the algorithm convergence slower and easier to overfitting, resulting in the model of the bounding box regression accuracy is not high. To solve this problem, this subsection improves it into the WIOUv2 bounding box regression loss function with dynamic focusing mechanism to improve the convergence process of the algorithm and enhance the detection accuracy of the algorithm.

WIOUv2 is improved on the basis of WIOUv1, so it is necessary to introduce WIOUv1. In the target detection training data, there are low-quality samples (the real frame and the predicted frame are very different), when generating high-quality predicted frames for low-quality samples, the commonly used geometric metrics make the algorithm penalize the low-quality samples too much, resulting in insufficient model generalization ability. In order to weaken the penalization strength of geometric factors, the distance metric is integrated into the distance attention to form the border regression loss function WIOUv1 with a two-layer attention mechanism. its expression is shown in Eq. (2); in order to explain R_{WIOU} , as shown in Fig. 3 the schematic representation of the smallest outer bounding rectangle containing the borders, corresponding to Eq. (3), (x, y) denotes the centroids of the prediction frames, and (x_{gt}, y_{gt}) denotes the center point of the real frame, W_g and H_g denote the width and height of the minimum outer rectangle, respectively, where the * marking in Eq. indicates that the part is stripped from the computational graph to eliminate its gradient gain.

$$Loss_{WIOUv1} = \mathrm{R}_{WIOU} Loss_{IOU} \tag{2}$$

$$\mathrm{R}_{WIOU} = \exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)^*}\right) \tag{3}$$

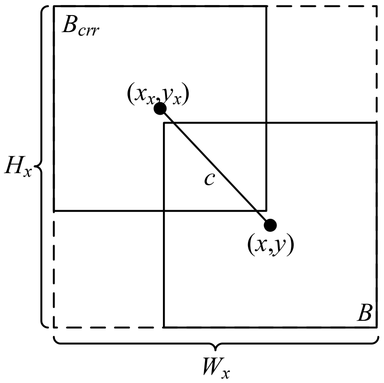


Figure 3: Schematic diagram with minimum bounding rectangle

Analyzing the expression $Loss_{WIOUv1}$, it can be seen that where $\mathrm{R}_{WIOU} \in [1,e]$, when generating ordinary quality prediction frames with smaller IOU values, the overall $Loss_{IOU}$ value is larger, and the overall $Loss_{IOU}$ of ordinary quality prediction frames will be further amplified under the influence of the factor; and $Loss_{IOU} \in [0,1]$, the distance metric R_{W10U} is smaller when a high-quality prediction frame is produced, which can further reduce the R_{WIOU} of the high-quality prediction frames due to the factor ranging from 0 to 1. The centroid distances tend to be 0 when the prediction frames are close to coinciding with the real frames. This reduces the focus of the loss function on the centroid distance.

On the basis of WIOUv1, drawing on the idea of FocalLoss focusing loss function, WIOUv2 newly introduces a monotonic focusing mechanism for cross entropy, which aims to reduce the proportion of simple samples contributing to the loss value, and at the same time, make the algorithm focus on the difficult samples to a larger extent, so as to improve the performance of algorithmic model localization. The WIOU-v2 expression is shown in Equation (4). Where $Loss_{IOU}^{\gamma^*}$ is the monotonic focusing coefficient, similarly, * indicates that this coefficient is also stripped from the computational map, $\gamma > 0$, $Loss_{IOU} \in [0,1]$, and therefore the focusing coefficient $Loss_{IOU}^{\gamma^*} \in [0,1]$·

$$Loss_{WIOUv2} = Loss_{IOU}^{\gamma^*} Loss_{WIOUv1}, \gamma > 0 \tag{4}$$

During model training, the monotonic focusing coefficient decreases as $Loss_{IOU}$ decreases, resulting in slower convergence in the later stages of training. To solve this problem, the mean value of $Loss_{IOU}$, $avg(Loss_{IOU})$, is used as the normalization factor, as shown in Eq. (5), and $avg(Loss_{IOU})$ is the mean value of the cross-comparison ratio loss of momentum m, whose value changes dynamically, which makes the model always have a high gradient gain during the training period, and solves the problem of slow convergence at the late stage of the model. The problem of slow convergence at the later stage is effectively solved.

$$Loss_{WIOUv2} = \left(\frac{Loss_{IOU}^*}{avg(Loss_{IOU})}\right)^{\gamma} Loss_{WIOUv1}, \gamma > 0 \tag{5}$$

II. C.UNet3+ network model

Although the improved YOLOv7 algorithm can accurately locate the damaged area, its output rectangular candidate box still contains a large amount of background noise, which is difficult to be used directly for refined analysis. For this reason, a semantic segmentation technique needs to be introduced to accurately categorize the pixels in the candidate box.

The encoder path consists of a convolutional layer and a downsampling layer, which halves the feature map and passes it to the next layer; the decoder path consists of an upsampling layer and a convolutional layer, which doubles the feature map and connects it to the jump-connected feature map. After the connection, UNet3+ reuses the features with dense blocks, and then passes the feature map through the convolutional layer to the output layer to output the predicted segmentation results, and the UNet3+ network structure is shown in Figure 4.

Advantages of UNet3+ network structure: through jump connection and dense block, local and global features of image can be captured at different scales to improve segmentation accuracy; and with few parameters and small memory occupation, it can work effectively under limited computational resources.

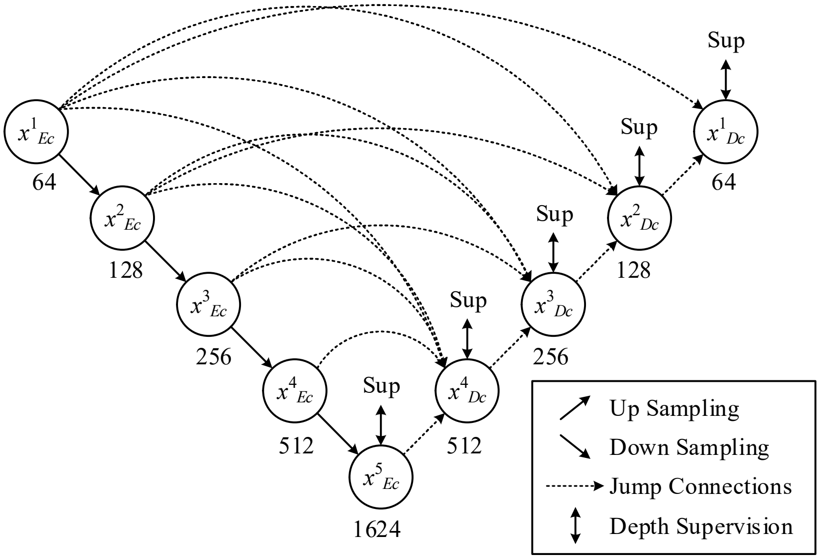


Figure 4: UNet3+ network structure

II. D.Damage Recognition Algorithm based on YOLOv7 with UNet3+

Although UNet3+ network can effectively segment cracked pixels, its individual application needs to traverse the whole map, and the computational efficiency is low. In order to balance the detection efficiency and segmentation accuracy, it is necessary to combine the fast localization ability of YOLOv7 with the fine segmentation advantage of UNet3+ to construct a two-level joint algorithm.

In the rectangular crack candidate box detected by YOLOv7, the area occupied by the damage is small, and the rest is background information. Threshold segmentation of this candidate target directly will introduce a large amount of noise due to the mis-segmentation of the background information, which is not conducive to the extraction of the subsequent geometric features of the crack. In order to eliminate the background information in this candidate box, this paper utilizes the image semantic segmentation network UNet3+ to perform pixel segmentation on the candidate box, i.e., classify it into crack pixels and background pixels pixel by pixel. The basic workflow: first the training set is labeled with the crack regions in each image. Secondly training is done using YOLOv7 for detecting the crack locations in the images. Training is performed using UNet3+, which is finally used for fine segmentation and recognition of the detected crack locations. The test set was tested using YOLOv7 to first locate the crack locations in the image and then UNet3+ was used for more accurate crack segmentation and recognition. The model is optimized and improved to increase the accuracy and performance of crack detection.

III. Detection and Experimental Analysis of Surface Damage of Ancient Buildings Based on Joint Detection-Segmentation Algorithm

Chapter 2 achieves the preliminary ability of ancient building material identification and damage localization through algorithm optimization and fusion, while in order to further validate the detection accuracy and robustness of the model in real scenarios, this chapter focuses on the specific experimental design, network training strategy and result analysis of surface damage detection.

III. A. Data sets

This paper takes the ancient architecture of “Gongshu Hall” as the research object, and the data set used for the experiment comes from the overall orthographic projection image of a section of the city wall with a pixel resolution of 55320×12320 provided by the Department of Ancient Architecture of “Gongshu Hall”. The images were manually divided into 6328 individual brick samples as the training set, which were divided into 4 categories, that is, 2371 "undamaged". "Flaked" 2016 pcs; 1072 "phthaline"; "Cracks" 869.

In the material classification task, the dataset is further subdivided into eight typical ancient building material classes: wood, brick, gray tile, stone, rammed earth, lime, glaze and metal.

In order to ensure that the number of samples in the training set is large enough, this paper adopts data enhancement operations, i.e., rotating, mirroring and changing contrast processing are implemented on the images, and the training set is finally expanded to 60,000 samples (15,000 for each category). At the same time, 4000 samples are manually segmented and categorized (1000 per class) at the position where the overall sample image is far away from the training set, which serves as the validation set. Finally, 1 partial wall image with 60 bricks is taken at the remaining locations of the wall sample for testing.

III. B. Ancient Architecture Material Recognition Training

Firstly, the improved EfficientNet v2 network is used to conduct training experiments for recognizing ancient building materials.

III. B. 1) Parameter setting

The EfficientNet v2 network, which introduces the mechanism of synergetic attention, will be developed under the Windows 10 operating system with the pytorch framework. pytorch is an open source artificial neural network library written in Python that provides a high-level neural network API, supports rapid experimentation, and is highly modular, concise, and extensible . Currently, pytorch is widely used in the field of deep learning and is considered a powerful and flexible tool.

In the model training process, the model is first constructed and trained using a training dataset, and the model parameters are optimized iteratively so that the model can more accurately predict the labels of a given input. The built model is then evaluated using a test set. In the process of optimizing the model, the model error on the test set is calculated and the model performance is measured using the cross-entropy loss function. By calculating the cross-entropy loss function, the difference between the model output and the actual labels can be measured, and the model can then be optimized to improve classification accuracy and generalization.

Hyperparameter optimization is a crucial step in deep learning model tuning, which can further improve the classification accuracy of the model. In the process of hyperparameter optimization, each major parameter of the model needs to be tuned, including the learning rate, the number of training rounds (epoch), the batch size (batch_size) and dropout.

(1) The learning rate is an important hyperparameter that controls the step size of the model to update the weights and determines the convergence speed and accuracy of the model. By adjusting the learning rate, the model can be made to converge faster, thus improving the accuracy of the model.

(2) The number of training rounds is the number of times the model is trained on the entire training dataset. By increasing the number of training rounds, the model can be made to learn the features in the dataset more fully.

(3) Batch size refers to the number of samples processed by the model during each training, and is also an important factor affecting the speed and stability of model training. By adjusting the batch size, the model can be made to better utilize the computational resources, thus improving the training efficiency and accuracy of the model.

(4) Dropout, which is an important regularization technique that can randomly set the output of a portion of neurons to zero, thus reducing the risk of model overfitting. By adjusting the value of dropout, the complexity and generalization ability of the model can be controlled.

In optimizing these hyperparameters, the grid search optimization method is used to find the optimal hyperparameter combination by training and evaluating different hyperparameter combinations, and each parameter is set as follows: learning rate: 1e-4; number of training rounds: 230; batch size: 32; dropout: 0.5, and the optimal hyperparameter combination further improves the model performance and generalization ability.

III. B. 2) Analysis of experimental results

The curves of loss function and accuracy change during the training process are shown in Fig. 5. From the curve, it can be observed that the training accuracy of the network gradually improves, and begins to stabilize at the training number of 230 times, and finally stabilizes at the level of 93.87%; at the same time, the value of the loss function gradually decreases, and begins to stabilize at the training number of 180 times, and finally stabilizes at 0.472.

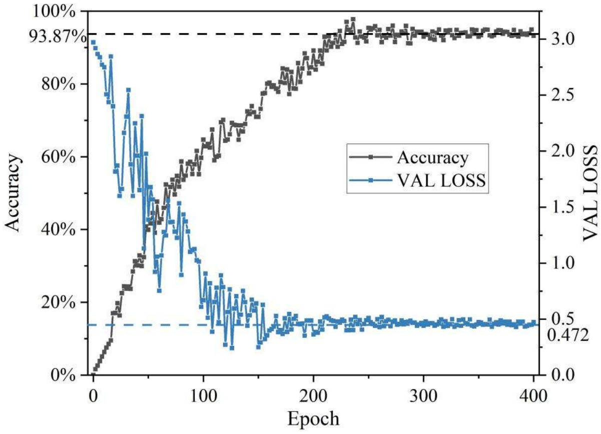
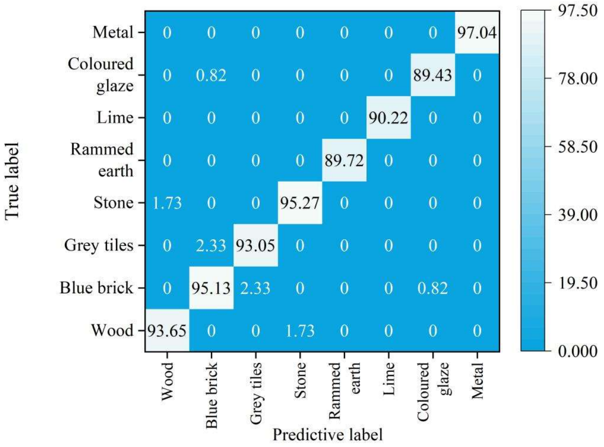


Figure 5: Loss function and accuracy rate variation

Next, various materials recognition of ancient buildings are trained, for the test sample confusion matrix of the ancient building material dataset is shown in Figure 6.



	Wood	Blue brick	Grey tiles	Stone	Rammed earth	Lime	Coloured glaze	Metal
Metal	0	0	0	0	0	0	0	97.04
Coloured glaze	0	0.82	0	0	0	0	89.43	0
Lime	0	0	0	0	0	90.22	0	0
Rammed earth	0	0	0	0	89.72	0	0	0
Stone	1.73	0	0	95.27	0	0	0	0
Grey tiles	0	2.33	93.05	0	0	0	0	0
Blue brick	0	95.13	2.33	0	0	0	0.82	0
Wood	93.65	0	0	1.73	0	0	0	0

Figure 6: The confusion matrix of test samples for ancient building materials

Figure 6 reflects the model's classification performance for each material. Overall, metal has the highest recognition accuracy of 97.04%, followed by 95.13% for green brick and 95.27% for stone, indicating that the model has a stronger feature extraction capability for materials with high contrast and regular texture. The main misclassifications are as follows: there is 1.73% cross misclassification between wood and stone, which may be due to the fact that the roughness of the texture of wood surface after weathering is similar to that of stone, resulting in feature confusion. The misclassification rate of green bricks and gray tiles is 2.33%, which may be related to the fact that they are both clay fired products with similar surface glaze reflective properties. Green bricks were also misjudged as glazed 0.82%, probably due to the fact that some of the green bricks were glazed on the surface and glazed with similar local reflective characteristics. The accuracy rates of rammed earth and lime are relatively low, 89.72% and 90.22% respectively, because their surfaces are susceptible to environmental erosion to form irregular mottling, which increases the difficulty of feature extraction. In addition, the accuracy rate of gray tile is 93.05%, and its misclassification is mainly focused on the 2.33% of green brick, which further confirms the similarity challenge of clay-based materials. The model performs weakly in glaze recognition with 89.43%, probably because its color diversity is not fully learned. Overall, the model performs well in most material classifications, but still needs to be optimized for materials with complex textures or susceptible to environmental influences.

III. C. Surface Damage Detection of Ancient Buildings

The optimization of material recognition provides the basis for damage detection, while the accurate detection of surface damage requires the combination of more complex network architecture and training strategy. In this section, we discuss the implementation and performance verification of damage detection based on the joint algorithm of YOLOv7 and UNet3+.

III. C. 1) Network training

The data needs to be preprocessed before network training, and in this paper we adopt zero-averaging. Zero-meanization, i.e., the average pixel value of all input images (training and validation sets are processed separately) is calculated, and then each pixel value is subtracted from the average value to obtain a new image instead of the original image as input, which reduces the effect of gradient inflation caused by a larger mean value. If some subsequent processing is required, such as principal component analysis, the data should have a zero mean. It should be noted that zero-meanization does not change the effective feature information of the image because it does not change the relative difference between pixels. For example, if an image of a brick with gray tones is transformed to white tones as a whole, it is still possible to determine whether it is damaged or not by the human eye, just as computer recognition of an image is not based on its absolute pixel values.

Network training using stochastic gradient descent + momentum method, the base of the same learning rate is set to 1e-4, the step size is 500, the decay rate is 0.87, that is, every iteration of 500 times will be multiplied by the learning rate of 0.87; training iteration 25 times for a validation, the maximum number of iterations is 5000; momentum coefficient is 0.89; weight regularization coefficient is taken as 0.0001. The variation of learning rate and verification accuracy with the number of iterations is shown in Fig. 7.

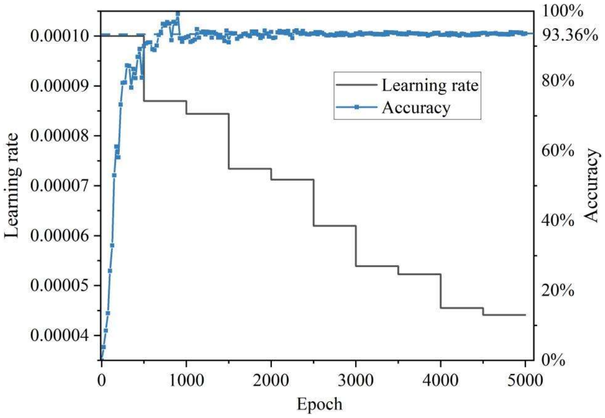


Figure 7: The learning rate and verification accuracy rate vary with iterations

III. C. 2) Analysis of loss measurement results

The optimization of network training strategy directly affects the final performance of the model. Through zero-mean preprocessing and dynamic learning rate adjustment, this section further analyzes the trend of the loss function and validation accuracy during the training process, and evaluates the model's effectiveness in detecting the three types of damage based on the experimental results.

For each type of brick sample, 10 test experiments are conducted, and the results are averaged, and the test results on the three types of damages, namely, “spalling”, “crumbling” and “cracking”, are shown in Figure 8.

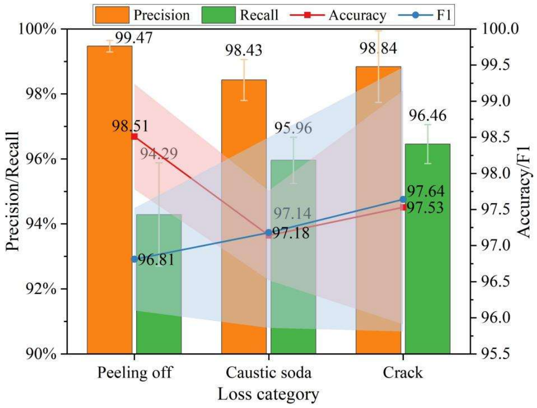


Figure 8: The test results of the three types of injuries

Figure 8 shows the performance of surface damage detection of ancient buildings based on the joint algorithm of YOLOv7 and UNet3, covering the accuracy, precision, recall and F1 values of three damage types: "spalling", "phthalate" and "crack". The experimental results show that the detection of flaking has the best overall performance, with an accuracy rate of 98.51% and a precision rate of 99.47%, indicating that the model has a very low false alarm rate for this type of damage; however, the recall rate is relatively low at 94.29%, which may be due to the fact that the boundary of some of the flaking regions is blurred, resulting in missed detection. The highest recall rate of 95.96% and an F1 value of 97.18% indicate that the model has a better control of miss-detection of crisps, but the precision rate of 98.43% is slightly lower than that of spallings, probably due to the similarity between the surface mottling characteristics of crisps and minor spallings. Cracks have the highest F1 value of 97.64%, and the recall rate of 96.46% is balanced with the precision rate of 98.84%, indicating that the model has a better ability to localize and segment cracks, thanks to the refined pixel-level segmentation strategy of UNet3+.

Overall, the model performs well in all three kinds of damage detection with an average F1 value of 97.21%, which verifies the effectiveness of the joint algorithm, but further optimization is needed for complex damage patterns such as edge blurring of spalling.

IV. Conclusion

In this study, for the digital preservation needs of the ancient architecture of “Gongshu Hall”, a set of intelligent media fusion technology framework based on image recognition and computer vision algorithms is proposed, and significant results are achieved in material recognition and damage detection. By improving the EfficientNet v2 network and introducing the collaborative attention (CA) mechanism, the model achieves a test accuracy of 93.87% on a dataset containing eight types of materials, of which metal (97.04%), brick (95.13%) and stone (95.27%) have the best recognition performance, while rammed earth (89.72%) and glazed glass (89.43%) have some misclassification due to the complexity of surface features, while rammed earth (89.72%) and glazed glass (89.43%) have some misclassification due to the complexity of surface features. The complexity of surface features has some misjudgment, which needs to further optimize the feature extraction strategy.

In terms of damage detection, the joint algorithm of WIOUv2 loss function and YOLOv7-UNet3+, which combines the dynamic focusing mechanism, significantly improves the localization and segmentation accuracy. The experimental results show that the average F1 value of the three types of surface layer damages (spalling, crisping, and cracking) reaches 97.21%, with the best performance in crack detection (F1=97.64%), and the precision rate of spalling (99.47%) and the recall rate of crisping (95.96%) are outstanding. The algorithm effectively reduces the background noise interference through the two-level detection-segmentation strategy, and provides reliable data support for geometric feature extraction and protection decision-making of ancient building damage.

Funding

The year 2023 Shaanxi Provincial Department of Education Scientific Research Program “This research was supported by the Research on digital protection and display of ‘Gongshu Hall’ based on intelligent media fusion technology” (23JK0261).

References

[1] Mohammed Mahmoud Mohammed Ahmed, O. (2019). New Approach for Digital Technologies Application in Heritage Architecture Conservation. International Journal of Multidisciplinary Studies in Architecture and Cultural Heritage, 3(1), 1-68.

[2] Fadli, F., & AlSaeed, M. (2019). Digitizing vanishing architectural heritage; The design and development of Qatar historic buildings information modeling [Q-HBIM] platform. Sustainability, 11(9), 2501.

[3] Wang, W., Yu, C. W., Peng, F., & Feng, Z. (2024). Digital development of architectural heritage under the trend of Metaverse: Challenges and opportunities. Indoor and Built Environment, 33(4), 603-607.

[4] Moreno Gata, K., & Echeverría Valiente, E. (2019). The use of digital tools for the preservation of architectural, artistic and cultural heritage, through three-dimensional scanning and digital manufacturing. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 501-506.

[5] Campi, M., Di Luggo, A., & Scandurra, S. (2017). 3D modeling for the knowledge of architectural heritage and virtual reconstruction of its historical memory. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 133-139.

[6] Koszewski, K. (2019, October). Visual representations in digital 3D modeling/simulation for architectural heritage. In Workshop on Research and Education in Urban History in the Age of Digital Libraries (pp. 87-105). Cham: Springer International Publishing.

[7] Liu, Q., Chen, J., & Chen, Y. (2018). A Research and Application of Three Dimensional Digitization Technology for Restore Chinese Architectural heritage from the Thirteen Factories. In MATEC Web of Conferences (Vol. 227, p. 02012). EDP Sciences.

[8] Agnello, F., Avella, F., & Agnello, S. (2019). Virtual reality for historical architecture. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 9-16.

[9] Nofal, E., Elhanafi, A., Hameeuw, H., & Vande Moere, A. (2018). Architectural contextualization of heritage museum artifacts using augmented reality. Studies in Digital Heritage, 2(1), 42-67.

[10] Galani, S., & Vosinakis, S. (2024). An augmented reality approach for communicating intangible and architectural heritage through digital characters and scale models. Personal and Ubiquitous Computing, 28(3), 471-490.

[11] Zhang, J. (2025). Spatial and temporal patterns and influential factors of ancient architectural heritage in Shanxi Province. Journal of Asian Architecture and Building Engineering, 1-22.

[12] Croce, V., Caroti, G., Piemonte, A., & Bevilacqua, M. G. (2021). From survey to semantic representation for Cultural Heritage: the 3D modeling of recurring architectural elements. ACTA IMEKO, 10(1), 98-108.

[13] Bayyati, A. M. (2017, July). Modern surveying technology: Availability and suitability for heritage building surveying and heritage building information models (HerBIM). In Heritage 2014-4th International Conference on Heritage and Sustainable Development.