International Journal of Gaming and Computer-Mediated Simulations
Volume 16 • Issue 1

Design of a MobilNetV2-Based Retrieval System for Traditional Cultural Artworks

Zhenjiang Cao, Shandong University of Arts, China*
Zhenhai Cao, Jingdezhen University, China

ABSTRACT

Aiming at the problem that it is difficult for art teachers to take into account each student in the art appreciation education in colleges and universities, this paper proposes a retrieval system for traditional cultural works of art. Dense connections are used to replace residual connections between bottlenecks in MobileNetV2 network and gradient transmission in the network. The dilution factor is used to control the size of the network to solve the problem of the rapid increase in the number of network channels. In addition, the non-local attention mechanism is effectively combined with the improved MobileNetV2 network structure, which effectively improves the classification accuracy of the network. Compared with VGG16, ResNet18, and ResNet34, the classification accuracy is increased by 21.3%, 9.2%, and 3%, respectively. The method in this paper has achieved good results in the classification of art works. According to the images of art works to be appreciated, it helps students understand the relevant cultural knowledge independently and reduce the burden of teachers.

KEYWORDS

Art Education in Higher Education, Attention Mechanisms, Deep Learning, MobilNetV2, Traditional Culture

1. INTRODUCTION

New media technologies, represented by computer vision, are increasingly becoming new variables that affect people's lives and work. As a new technology leading to changes in people's lives and work, artificial intelligence has the same disruptive potential in the work of education and teaching. In current Chinese university education, art education is responsible for enhancing students' aesthetic skills and passing on traditional culture, an area in which science and technology students are undoubtedly lacking compared to those majoring in art (Feng et al, 2021). Art majors have systematic professional art teaching, usually in small classes, and teachers can generally take care of every student. However, as students of science and technology, due to the different training requirements, art education in schools is usually given in the form of elective courses, mainly in the form of appreciation of art works, and the excessive number of elective courses makes it difficult to ensure that each student has a detailed understanding of the works to be appreciated in class. Utilizing computer vision technology

DOI: 10.4018/IJGCMS.334700 *Corresponding Author

to categorize artworks featuring traditional culture holds immense significance in imparting associated cultural knowledge within college art education. Nonetheless, artworks embedding traditional cultural elements often feature diminutive representations of the cultural essence within the image, thereby heightening the challenge for network recognition.

Neural networks exhibit commendable performance in tackling nonlinear conundrums, exemplified by their proficiency in tasks like image recognition. However, they are not without their design challenges, including the intricacies of determining network architecture, the selection of training parameters, and the delicate calibration of network weights. These quandaries necessitate substantial human involvement in the neural network design process. Within the realm of neural network structure exploration, the parameters constituting the search space play a pivotal role in defining the network's structure (Chen et al, 2023). This encompasses critical aspects such as the number of network layers, the dimensions of convolutional kernels, and the choice of operators for each layer. The quest for an optimal network structure has emerged as a focal point in deep learning research in recent years. By employing efficient and cost-effective search algorithms, it becomes possible to automatically ascertain a neural network structure endowed with robust generalization capabilities and hardware-friendly attributes. This approach not only obviates the need for manual network design but also outperforms early, labor-intensive design paradigms across a spectrum of application scenarios (Zhong et al, 2023).

Feature extraction from images constitutes a pivotal phase in the realm of traditional image classification techniques. Researchers invariably find themselves in the position of crafting custom feature extraction methods tailored to the unique characteristics of the images within a specific classification task. A case in point is the Scale-Invariant Feature Transform (SIFT) (Kasiselvanathan, 2020), which excels in discerning object features for the purpose of object recognition. SIFT leverages the presence of salient points in an image, regardless of whether they are distant or nearby, blurred or sharp. Similarly, the Local Binary Pattern (LBP) operator (Cheng et al, 2022) is a stalwart in extracting edge information from facial images, primarily for face recognition. LBP capitalizes on the relationships between pixel gradients and their immediate neighbors to uncover valuable facial features.

For pedestrian detection, the Histogram of Oriented Gradients (HOG) feature (Yang & Wei, 2022) becomes the method of choice. HOG operates by capturing alterations in directional gradients at the juncture of the target object and its background. This approach effectively teases out the edge gradient features of the human body in images, making it indispensable for pedestrian detection tasks.

Besides the conventional image classification methods mentioned above, numerous alternative methods have delivered outstanding results in diverse image classification tasks. However, one common challenge with these methods is the necessity for manual parameter configuration during feature extraction. This reliance on manual intervention presents a formidable drawback as it does not guarantee consistent classification performance when transitioning from one image to another. In the backdrop of progressively intricate image classification tasks, it becomes increasingly evident that methods hinging on manual parameter tuning are no longer equipped to meet the demands of practical applications (Xu et al, 2023). As such, there is a pressing need for the development of more automated and adaptive feature extraction techniques to ensure robust and reliable classification results across various image datasets.

The advent of convolutional neural networks (CNNs) has ushered in a revolutionary era of remarkable advancements in image classification. This transformative progress can be attributed to the rapid evolution of deep learning, which has injected fresh vitality into the increasingly intricate landscape of image classification tasks. For instance, the VGG architecture, as exemplified by VGG(Thepade et al, 2022), leverages small convolutional kernels to significantly deepen the network. This strategic design has led to substantial improvements in performance. By virtue of this architectural innovation, VGG has not only expanded the network's capacity to extract intricate features but has also elevated its overall classification accuracy. In a similar vein, ResNet(Zhao et al, 2022) and DenseNet(Albahli et al, 2021) have introduced groundbreaking concepts in neural network

design. ResNet exploits residual connectivity, allowing the network to perform tens of thousands of iterations without succumbing to vanishing gradient problems. This enables the model to achieve image recognition capabilities that surpass even the acumen of human observers. DenseNet, on the other hand, employs dense connectivity, fostering a tightly woven network structure that encourages feature reuse and information flow across layers. This not only enhances classification accuracy but also ensures robust training.

However, an inherent challenge emerges with the remarkable accuracy attained by these advanced CNNs. The proliferation of network parameters, necessary to achieve such accuracy, results in models of substantial size. These parameter-heavy models, while excelling in performance, pose a practical challenge when it comes to deployment on resource-constrained devices such as mobile phones and tablets. The computational demands of these models can overwhelm the processing capabilities and memory limitations of these devices, making them less suitable for real-time, on-device image classification tasks. Consequently, there is a pressing need to explore model compression and optimization techniques to make these powerful networks more accessible and practical for a wider range of applications and devices.

Enhancing the precision in classifying traditional cultural elements within artworks remains pivotal, enabling students to access comprehensive information about these pieces through their mobile devices or tablets during classroom appreciation sessions. In this paper, a MobileNetV2 classification network with Non-Local attention mechanism is proposed and experimented on the collected image dataset containing traditional cultural artworks, which can well identify the traditional cultural elements in the images. The primary contributions encompass:

- Utilizing dense connections in lieu of residual connections within MobileNetV2's bottlenecks, coupled with feature reuse techniques, optimizing information and gradient propagation efficiency within the network.
- Introducing a dilution factor aimed at regulating the MobileNetV2 network size, mitigating the issue of rapid expansion in channel numbers.
- Effectively integrating the non-local attention mechanism into the enhanced MobileNetV2 network structure, thereby significantly enhancing the network's classification accuracy.

2. RELATED WORK

The classification of artworks containing traditional cultural elements is an important area of research. In university art education, deep learning techniques can help students to filter traditional cultural elements in images and reduce the pressure faced by teachers in large classes, and more and more researchers are using computer vision techniques to classify artworks containing traditional cultural elements.

Ever since LeCun et al. proposed the LeNet-5 model (Sun et al, 2021), which used stacked convolutional and pooling layers to extract features from input handwritten character images, allowing a significant improvement in the classification accuracy of handwritten characters, the convolutional neural network approach has brought a new light to the bottlenecked image classification task. AlexNet (Gu et al, 2022) added three layers to LeNet-5 AlexNet added three convolutional layers to LeNet-5, turning it into a deeper and larger model, and achieving classification performance that was not possible with traditional methods. Later VGG increased the depth of the network to 19 layers, further improving the classification ability of the network. However, the brutal deepening of the network layers made it difficult to train the network, and the image classification task was again bottlenecked. DenseNet incorporates the idea of feature reuse from ResNet and proposes a dense connection that connects each layer, further enhancing the network's ability to dig deeper into image information. From LeNet-5 to DenseNet, the deepening of the number of layers in the network has led to a rapid increase in classification performance, but also to an increase in the number of parameters. Although

the rapid development of hardware has brought more computing power, mobile devices such as mobile phones and tablets are still unable to handle such large models. To solve the problem that the model is too large to be used on mobile devices, MobileNetV1 (Zhao et al, 2021) divides the entire feature map in the original convolutional operation into different channels for convolutional operation, which greatly reduces the number of parameters while ensuring the performance is comparable to that of normal convolution. MobileNetV2 (Hussain et al, 2022), like DenseNet, introduces the idea of ResNet feature reuse, but due to the smaller number of parameters in the network, the performance improvement of MobilNetV1 is achieved by using the opposite lift-dimensional operation in the Block of MobilNetV2 as that of ResNet.

However, for artworks with traditional cultural elements, sometimes the small size of the targets in the image that represent traditional culture makes it more difficult for the network to recognize them. Some recent studies have used human-specific visual mechanisms to allow the network to progressively deepen its focus on these smaller sized traditional cultural elements during the training process. Non-Local (Tuna et al, 2022) networks focus on the information that is important in the whole feature map by calculating the similarity between pixels. This method has effectively improved the network's ability to capture features, and several researchers have since begun to work on this method of capturing features. DANet (Fu et al, 2020) enhances the features to be focused on by considering the degree of relevance of each location on the feature map using Non-Local learning and the dependencies between channel mappings. ANLNet (Zhu et al, 2019) samples the key and $value$ obtained from high-level features using a pyramidal pooling of the null space with four different expansion rates, respectively, and the low-level features are convolved and used as $query$, effectively reducing the cost of collecting global contextual information for the network. Also to reduce the cost of collecting contextual information from the network, CCNet (Huang et al, 2020) focuses only on the relationship between pixels on the feature map and pixels in the same row and column during one calculation of attention, and obtains the relationship for the full image element by two calculations, obtaining better results while effectively reducing the computational cost.

MobileNetV2 is a lightweight convolutional neural network model proposed by Google in 2018 (Xiang, et al., 2019). Li et al. (2023) used MobilenetV2 to realize the rapid identification of corn grain images, and the final model accuracy reached 97.95%. Xu et al. (2020) built MobileNetV2 network model to classify domestic waste images. Experimental results show that the classification accuracy of the improved MobileNetV2 network model on the domestic waste classification dataset is 90.58%. In other studies (Wang et al., 2022; Pavithra et a., 2023), experiments show that the lightweight MobileNetV2 not only has low parameters, but also performs well in multi-classification tasks.

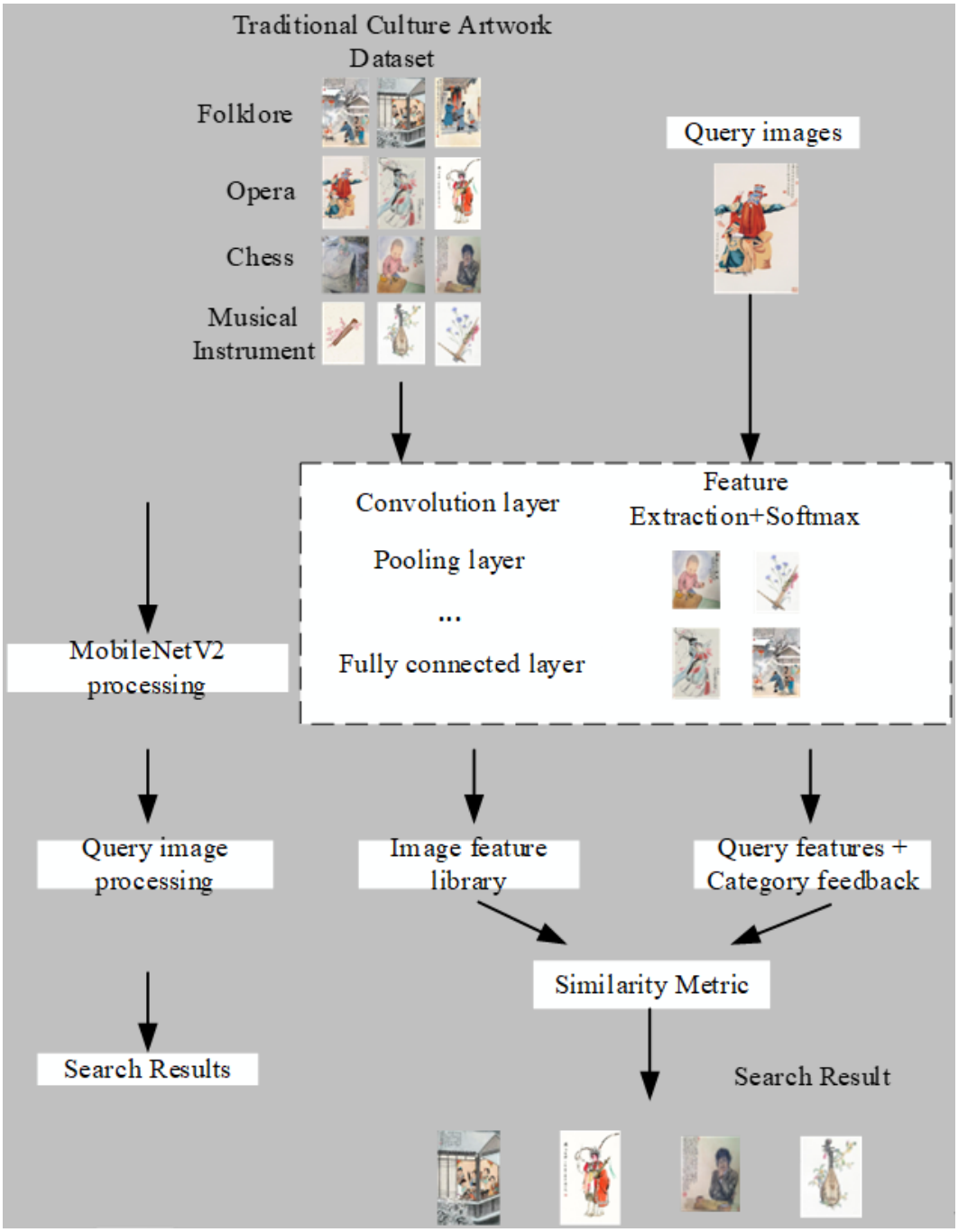
3. IMPROVED SEARCH SYSTEM FOR TRADITIONAL CULTURAL ARTWORKS

The processing flow of the proposed traditional culture-based artwork retrieval system is shown in Figure 1. First, the MobileNetV2 network with Non-local attention mechanism is properly trained to achieve parameter learning of the network and obtain Softmax classifiers that meet the requirements. Then the feature extraction operation is performed according to the requirements, and the corresponding image features are obtained, based on which the feature library is further determined. After processing the features and categories of the input image, the similarity measure is performed with the features of the image library, and the corresponding retrieval results are determined after sorting, and the corresponding process is as follows.

3.1 Inverse Residual Connectivity in MobileNetV2 Network

The high computational resources required by today's convolutional neural networks far exceed the carrying capacity of edge computing devices. mobilenetv2 network is a new network structure designed to address this problem. The network can effectively reduce model parameters and computation while maintaining high detection accuracy. MobileNetV2 stands out due to its innovative approaches,

Figure 1. Search processing flow chart



particularly the utilization of the linear bottleneck and inverse residual connectivity. These elements revolutionize the network's architecture to optimize performance. The linear bottleneck strategy involves removing layers with smaller output dimensions, thereby minimizing feature loss that

typically occurs with the Relu activation function. By transitioning to a linear activation function, MobileNetV2 effectively reduces feature loss, enhancing the network's overall efficiency.

The inverted residual connection is another key feature that distinguishes MobileNetV2. This design involves lifting dimensions before reducing them, a departure from the traditional residual module. This inversion significantly enhances the network's ability to reuse features, thereby improving its overall performance. A shift to dense connections between bottlenecks in the MobileNetV2 network is proposed instead of relying on residual connections. This change aims to leverage feature reuse techniques, ultimately enhancing information and gradient transmission efficiency within the network. However, it's crucial to note that while dense connections utilize splicing, unlike summation used in residual connections, this alteration leads to a rapid increase in the number of bottleneck output channels. Consequently, this augmentation escalates the network's parameter count and computational time consumption. y adopting these groundbreaking strategies, MobileNetV2 achieves significant advancements in feature preservation, network efficiency, and performance optimization, establishing itself as a prominent model in the field of mobile neural networks

Therefore, in this paper, a dilation factor t that can be dynamically sized between 1 and 6 is designed to control the size of the network. A bottleneck with a dilation factor of 1 is used in designing the network, but the original 1×1 convolutional layer is not removed in designing the bottleneck because 1×1 convolutional not only improves the performance of the network but more importantly, the 1×1 convolutional layer can linearly combine multiple feature mappings, thus enabling the integration of features across channels. Since the network introduces dense connectivity, it can be assumed that even if the initial value of the expansion factor in the initial 1×1 convolution is set to 1, it will have a positive impact on the improvement of the network performance. The bottleneck structure of the MobileNetV2 network designed in this paper is shown in Figure 2. The inputs and outputs of each of these layers are shown in Table 1. k_{in} denotes the number of input channels, h denotes the height of the input, w denotes the width of the input, t denotes the dilation factor, s denotes the step size, and k_{out} denotes the number of output channels.

3.2 Non-Local Attention Mechanism

In the rapidly evolving landscape of deep learning architectures, the fusion of novel attention mechanisms with established network structures has sparked remarkable advancements in image classification. One such integration lies in the marriage of the Non-local attention mechanism with the robust MobileNetV2 architecture. This synergy aims not only to harness the efficiency and proven performance of MobileNetV2 but also to augment its capabilities by incorporating Non-local attention, renowned for its proficiency in capturing long-range dependencies within images. The amalgamation

Figure 2. Improved bottleneck structure

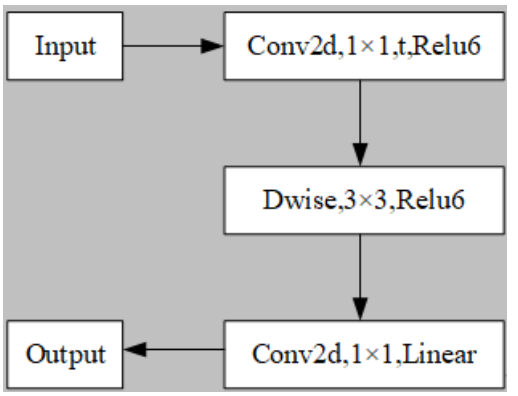


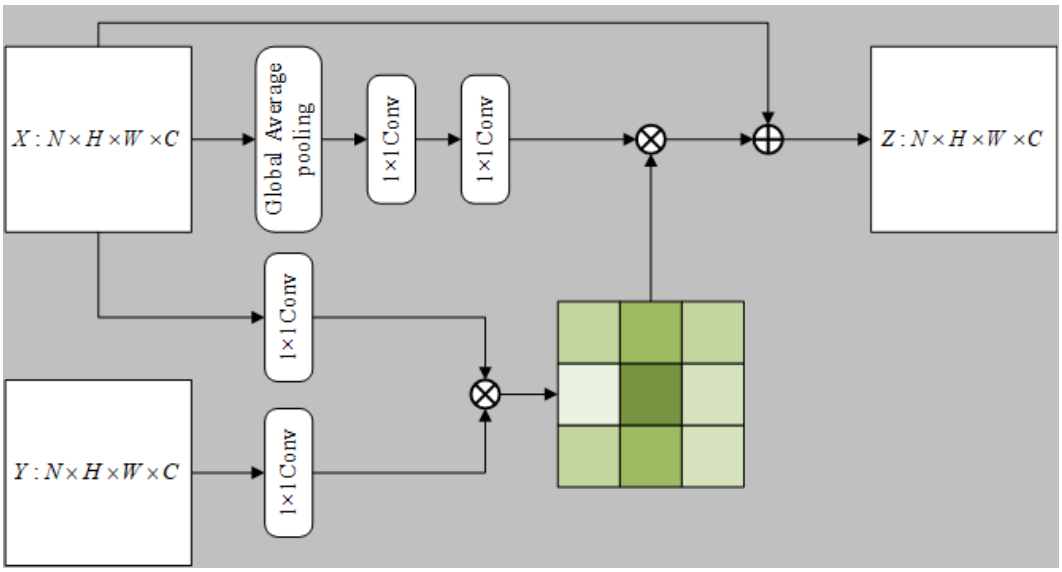
Table 1. Input and output of bottleneck layers

Input	Operation	Output
$h \times w \times k_{in}$	$1 \times 1 Conv2d, \text{Relu}\,6$	$h \times w \times \left(tk_{in}\right)$
$h \times w \times \left(tk_{in}\right)$	$3 \times 3 Dwise$ $s = s, \text{Relu}\,6$	$\frac{h}{s} \times \frac{w}{s} \times \left(tk_{in}\right)$
$\frac{h}{s} \times \frac{w}{s} \times \left(tk_{in}\right)$	$Linear, 1 \times 1 Conv2d$	$\frac{h}{s} \times \frac{w}{s} \times k_{out}$

of these techniques holds the promise of significantly elevating the classification accuracy of the network while addressing complex relationships among diverse features.

Non-local attention mechanism can ignore irrelevant information in the feature map and focus on point information, extract more semantic information, are computationally efficient, and are easy to be embedded in various network structures. Therefore, in this paper, the Non-local attention mechanism is effectively integrated with the improved MobileNetV2 network structure to effectively improve the classification accuracy of the network. As shown in Figure 3, the Non-local attention mechanism takes the feature maps X and Y as input, and first adds a global average pooling layer with 1×1 convolution before X to obtain global feature information, and then performs linear mapping of X and Y to obtain Q, K, V features; through Reshape operation, the dimensions of the above three features except the number of channels are combined, and then performs matrix dot product on Q and K The self-correlation in the features is calculated by the Reshape operation, and the relationship of each pixel in each frame to all pixels in all other frames is obtained; the Softmax operation is performed on the self-correlation features to obtain the weights with the value range of [0-1], i.e., the required Self-Attention coefficients; finally, the Self-Attention coefficients are multiplied

Figure 3. Cross-layer implicit feature pyramid network



back into the feature matrix V, and then the 1×1 convolution is performed with the original then the 1×1 convolution is done with the original input feature map A for residual operation to obtain the output Z.

The improved MobileNetV2 network bottleneck with a Non-local attention mechanism is introduced into MobileNetV2 to improve the real-time and accuracy of classification. The specific structure of the network is shown in Figure 4. Conv_1 to Conv_13 are the 13 convolutional layers in MobileNetV2. In these 13 convolutional layers, deep convolution and point-by-point convolution are used to extract features from each other. Among them, Conv_1, Conv_3, Conv_5, Conv_7-11, and Conv_13 use deep convolution - point-by-point convolution to avoid the problem that deep convolution only extracts features in low-dimensional space. Conv_2, Conv_4, Conv_6, and Conv_12 convolution layers use point-by-point convolution - deep convolution - point-by-point convolution to avoid the problem that the Relu6 activation function in deep convolution will destroy the feature extraction. The Conv_4, Conv_6, and Conv_12 convolutional layers are selected to implement the Non-local attention mechanism from deep to shallow upsampling and lateral connectivity.

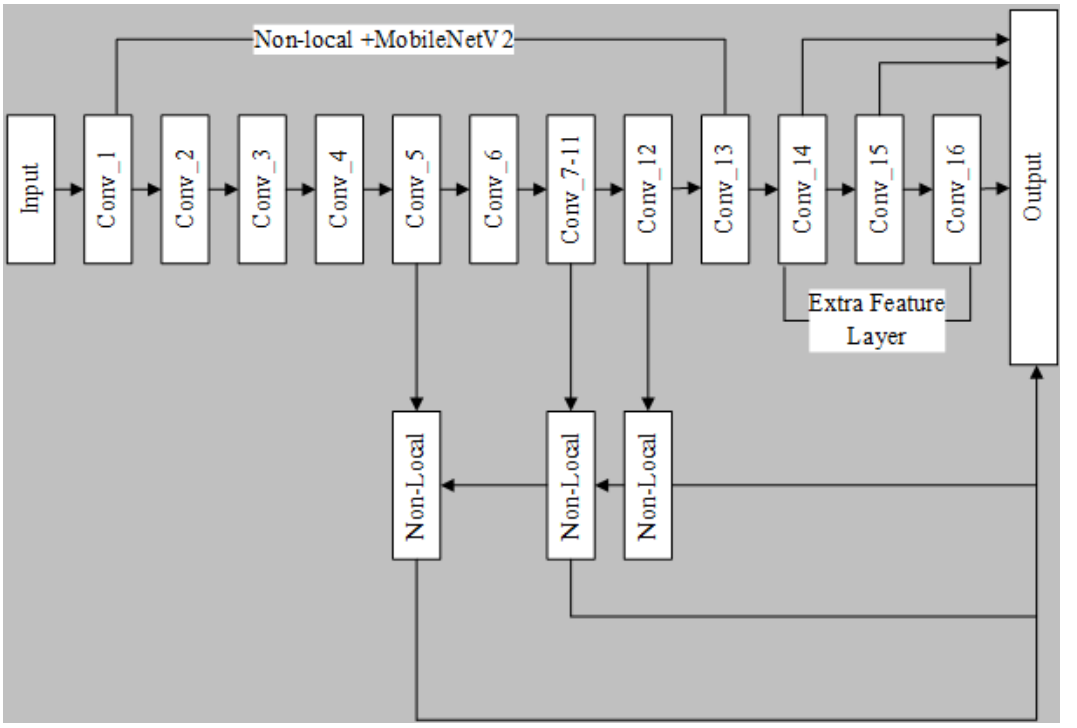
3.3 Improvement of Loss Function

Lin et al. (2017) proposed the focal loss function and the expression is shown in the following equation.

$$FL\left(p_t\right) = -\left(1-p_t\right)^{\gamma} \log\left(p_t\right), \tag{1}$$

where p_t is the classification probability of different categories. focal loss controls the weight of simple samples through the modulation factor $\left(1-p_t\right)^{\gamma}$. This factor reduces the weight of samples

Figure 4. Network structure



that are easy to classify during the training period and therefore prefers to train samples that are difficult to classify when the model is trained during the training period.

Although focal loss can solve the problem of positive and negative sample imbalance in the training process, some classification scores are low but the results are true, and some classification scores are high but the results are false, which will affect the accuracy of image classification in the training process of the model. To solve this problem, Li et al. (2020) improved focal loss and proposed Quality focal loss. in this paper, with the idea of the quality focal loss function, Quality focal loss is introduced into the MobileNetV2 network to adjust the weights of simple samples and difficult samples, so that the model focuses on the training of difficult samples.

The definition of quality focal loss is shown in the following equation.

$$QFL\left(\ell\right) = -\left|y-\ell\right|^{\beta}\left(\left(1-y\right)\log\left(1-\ell\right)+y\log\left(\ell\right)\right), \tag{2}$$

Quality focal loss extends focal loss in two main steps.

1. extend the cross-entropy part $-\log\left(p_t\right)$ to $-\left(\left(1-y\right)\log\left(1-\ell\right)+y\log\left(\ell\right)\right)$.
2. expand the modulation coefficient $\left(1-p_t\right)^{\gamma}$ into $\left|y-\ell\right|^{\beta}\left(\beta\geq 0\right)$, ℓ is the output of the sigmoid function, and the parameter β controls the weight between positive and negative samples.

Experimentally, it is proved that the easier the samples are balanced and the loss is smaller when $\beta = 2$, thus improving the classification performance of the network.

3.4 Similarity Metric

In this analysis this paper measures the similarity between two fine art images based on Euclidean distance. x and y correspond to the feature vectors of both, and the corresponding Euclidean distance expressions are as follows.

$$D\left(x,y\right) = \left(\sum_{i=1}^{n}|| x_i - y_i ||^2\right)^{1/2} \tag{3}$$

For the queried artworks, the relevant image features are determined by processing based on the trained convolutional neural network, while the Softmax classifier is used for classification processing, and the image category is determined, then the feature maps of the images of the same category are retrieved and the Euclidean distance between the two is calculated to achieve image retrieval. The retrieval results are returned in a certain order according to the size of the feature distance results. After the retrieval is completed, the system will display the category of the artwork and the related cultural knowledge on its own.

3.5 GA Search

Applying linear phase constraints to convolution filters within a network can indeed lead to a notable decrease in the number of network parameters. However, this efficiency gain often accompanies a trade-off: a considerable decline in network accuracy. To strike a balance, a compromise is made by subjecting only a portion of the filters within the network to these linear phase constraints.

In the genetic algorithm implemented, the chromosome structure is designed to handle this nuanced approach. Each chromosome consists of 39 binary digits, aligning with the 13 convolutional layers in the MobileNet structure. Notably, every third gene point in the chromosome encodes the

proportion of symmetrical filters with linear phase constraints allocated to each layer. This arrangement enables fine-tuning the application of these constraints across the network in a controlled and strategic manner. The procedure unfolds as follows:

(1) Encoding and decoding: The search parameters in this paper are the filter ratio examples with linear phase reduced beam contained in each layer of the 13 1×1 convolutional layers of the MobileNet network.
(2) Population Initialization: Initial genes of individuals are randomly generated using a random number function. The population comprises 400 individuals, and the evolutionary process spans 500 generations.
(3) Population Assessment: The fitness of each individual is determined based on the accuracy achieved under various network structures with different proportions of symmetric filters featuring linear phase constraints. This fitness value guides the direction of the genetic algorithm.
(4) Selection: The roulette wheel algorithm is employed for individual selection, proportionally inheriting individuals to the next generation population based on their fitness.
(5) Crossover Operation: A single-point crossover method is utilized to exchange a gene at a specific point on the chromosome of each individual, generating new chromosomes.
(6) Mutation Operation: Employing binary coding, mutation is enacted by inverting specific chromosome points. In this paper, single-point mutation is the chosen method.
(7) Termination Criteria: When the genetic algorithm evolves over a certain number of generations and the majority of individual genes converge, it is indicative of searching the global optimal solution for the objective function. At this juncture, if the population ceases to evolve within a fixed number of generations, it is considered that the genetic algorithm has identified the optimal MobileNet network structure with filters bearing the requisite linear phase constraints, as stipulated in this study.

The parameters set for the Genetic Algorithm (GA) play a pivotal role in determining its efficiency and effectiveness in optimizing solutions. In this specific scenario:

Number of Individuals: 100 individuals constitute the population within the GA. This population size affects diversity and exploration-exploitation trade-offs in the search space. A larger population may enhance exploration but might also increase computational overhead.

Genetic Generations: The GA undergoes 100 generations, implying that the evolution process (selection, crossover, mutation) is repeated for 100 iterations. The number of generations influences convergence and the search for optimal solutions.

Crossover Probability: The probability of crossover between individuals during reproduction is set at 0.75. This value determines the likelihood of genetic material exchange between selected individuals, promoting diversity and convergence in the population.

Mutation Probability: The mutation probability stands at 0.05, indicating the likelihood of random changes occurring in the genetic material of individuals. Mutation introduces diversity and prevents premature convergence, aiding in exploration of the solution space.

These parameter settings represent a balance between exploration and exploitation within the GA framework. The chosen values seem to prioritize diversity and exploration (larger population, moderate crossover probability, and a relatively low mutation probability), which can be beneficial for discovering diverse solutions while avoiding premature convergence to suboptimal solutions.

This approach acknowledges the delicate balance between parameter reduction and network accuracy, allowing for a tailored application of linear phase constraints within the MobileNet architecture. By optimizing the distribution of these constraints across layers via the genetic algorithm, it aims to harness the advantages of parameter reduction while mitigating the adverse impact on overall network performance.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The experiments are conducted in Python on Windows 10, using the GPU version of the deep learning framework TensorFlow 1.4.0, which can be accelerated with the GPU provided by CUDA when using the CPU for training. The processor is Intel i7-10875H, the memory is 15GB, and the graphics card is NVIDIA GeForce RTX 2070. The Batch_size is set to 128, the momentum is set to 0.9, and the initial learning rate is set to 0.01. The model is trained for a total of 200 iterations, and the learning rate is changed to 0.001 after 137 iterations.

Efficiency and optimization are key considerations in the design of the MobileNet network, particularly in relation to the distribution of 1×1 convolutional filters with linear phase constraints within each layer. In this paper, to streamline the exploration process, the search space is deliberately constrained. Initially, predetermined ratios of filters adhering to linear phase constraints are assigned to each layer of the MobileNet network, specifically set at 30%, 50%, and 70%. This targeted approach aims to expedite the search for an optimal network structure while maintaining precision and performance.

The exploration process then unfolds within the confines of these preset ratio constraints. The objective is to identify the most effective network configuration within these limitations. Once this optimal structure is discerned, the design incorporates an additional proportion of filters with linear phase constraints into the network architecture. This expansion of filters further bolsters the model's precision and performance, culminating in an even more efficient and effective network.

The comprehensive evaluation of the proposed system's efficacy involved the curation of a dataset comprising 1,105 traditional cultural art images sourced from diverse domestic art platforms, microblogs, and Zhihu. This initial dataset underwent augmentation techniques, expanding it to a substantial 4,420 images. The augmentation process aimed to diversify and enrich the dataset, enhancing its variability and inclusivity.

This augmented dataset was meticulously categorized into various image types, with a detailed breakdown presented in Table 2. The thoughtful classification and distribution of images across different categories ensured a balanced representation, facilitating a robust evaluation of the designed system's performance.

The diversity encapsulated within this dataset not only reflects the richness of traditional cultural art but also serves as a comprehensive test bed for the proposed system's capabilities in the realm of traditional cultural art image classification and analysis. Its expansive nature enables the system to encounter a wide array of scenarios, enhancing its adaptability and ensuring a more thorough assessment of its performance across various art forms and styles.

By leveraging this extensive and diverse dataset, the evaluation process aimed to provide a comprehensive understanding of the system's strengths and limitations, ultimately validating its efficiency and suitability for the nuanced task of traditional cultural art image analysis and classification.

Table 2. Classification of fine art image categories and number of images in each category

Fine Art Image Categories	Training Set	Test Set	Total
Folklore	400	390	1290
Opera	1130	400	1530
Chess	660	240	900
Musical Instrument	510	190	700

4.1 Comparison of Search Results

There are many different metrics to evaluate the performance of image retrieval approaches, and in this paper, Capability of Precision (CP) and recall Precision (RP) are used to evaluate the processing retrieval results during the study, and the corresponding expressions.

$$CP = NTP/NP \tag{5}$$

$$RP = NRTP/NR \tag{6}$$

Where, NTP is the number of correctly classified images in the test image, NP is the total number of images in the test image library, NRTP is the number of images related to the query image in the retrieval result, and NR is the total number of images returned by the retrieval result.

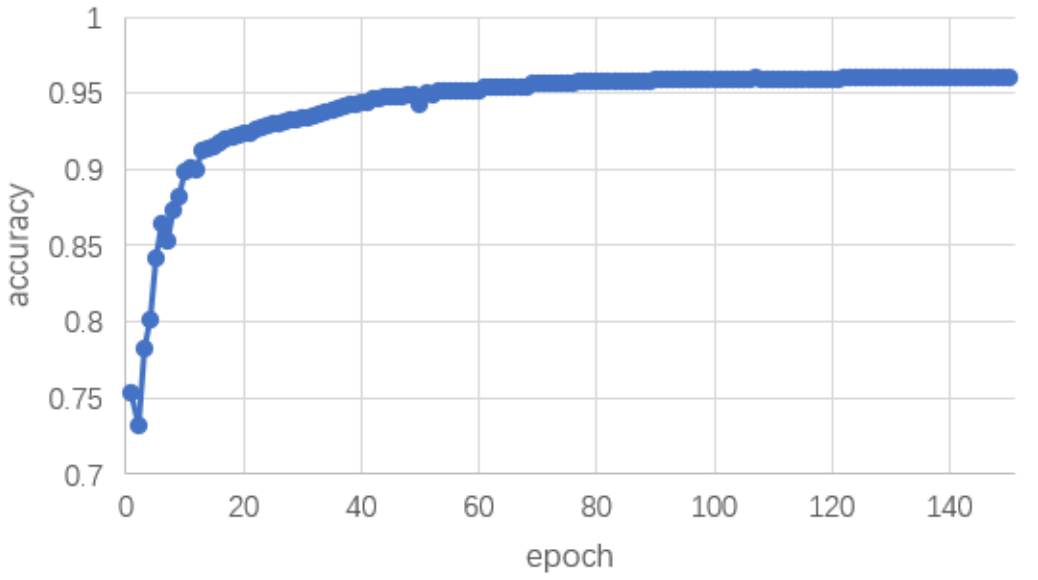
The network related to Figure 5 was used to train, and after 150 iterations of operation, it was found that the classification accuracy of the network was over 96%. The corresponding classification accuracy curves are shown in Figure 5.

Figure 5 offers a noteworthy insight into the performance of the MobileNetV2 network enhanced with the Non-local attention mechanism as proposed in this paper. The plot reveals a distinctive behavior during different iterations. In the initial phase, spanning from 0 to 20 iterations, the accuracy curve exhibits an upward trajectory, albeit with noticeable oscillations. This oscillatory behavior can be attributed to the network's adaptation and learning process in its early stages.

However, from iteration 20 through 150, the accuracy curve attains a remarkable degree of stability. This plateau signifies that the proposed model showcased in this paper achieves convergence with consistent and reliable performance. The network appears to have settled into a state where its accuracy remains relatively constant, reflecting a strong convergence speed and a high degree of generalization ability.

Throughout the study, a subset of 20 images was randomly selected from each of the four categories within the dataset for testing purposes. The study further expanded its scope by analyzing

Figure 5. Classification accuracy curve



the retrieval accuracy concerning different retrieval settings. Specifically, the number of retrieved images was systematically set to 20, 40, 60, 80, and 100, and the corresponding retrieval accuracy was meticulously computed and compared based on these settings. The outcomes of this analysis, along with the detailed results, are thoughtfully presented in Table 3.

This extensive evaluation of retrieval accuracy under various settings reinforces the validity and practical utility of the proposed MobileNetV2 network with the Non-local attention mechanism. It attests to the model's consistent performance and robustness in real-world retrieval scenarios, highlighting its potential for diverse fine art image retrieval applications.

4.2 Model Analysis

The architecture of CNN plays a pivotal role in shaping the accuracy of classification results, which, in turn, impacts the accuracy of retrieval outcomes. Therefore, in the course of this study, an in-depth investigation into classification performance was undertaken, employing three distinct network architectures. The results were meticulously analyzed and compared to shed light on the relationship between the network structures and the classification accuracy under the respective conditions. The findings of this comparative analysis are succinctly summarized in Table 4.

This comprehensive exploration of various network architectures serves to underscore the critical importance of selecting an appropriate model for fine-tuning classification and retrieval tasks. The data presented in Table 4 reveals the tangible impact of the network's structural choices on classification accuracy, providing valuable insights for researchers and practitioners in the field of fine art image analysis and retrieval.

The results presented in Table 4 underscore the remarkable performance of the network proposed in this paper, surpassing renowned models such as VGG16, ResNet18, and ResNet34 in both training time and classification accuracy. This notable achievement can be attributed to the substantial enhancements made to the MobileNetV2 architecture within this study, leading to a significant reduction in network complexity and computational overhead.

One key modification involves the removal of residual connections between MobileNetV2 bottlenecks. This strategic alteration significantly amplifies the reuse rate and transfer efficiency of fine art image features. Consequently, the network demonstrates a remarkable 21.3% boost in

Table 3. Fine art images dataset retrieval accuracy results

Number of Images	Folklore	Opera	Chess	Musical Instrument
20	1	1	1	1
40	1	0.998	1	1
60	0.995	0.992	0.991	0.994
80	0.989	0.988	0.986	0.987
100	0.984	0.982	0.981	0.980

Table 4. Network knot model classification accuracy comparison chart

Network Structure	Training Time/h	Classification Accuracy	Number of Network Layers	Network Size
Vgg16	2.725	0.751	16	42MB
ResNet18	2.124	0.872	18	49MB
ResNet34	2.482	0.934	34	60MB
Ours	1.989	0.964	13	19MB

classification accuracy compared to VGG16, a 9.2% improvement over ResNet18, and a 3% increase when compared to ResNet34. These gains in accuracy reflect the superior capacity of the modified MobileNetV2 architecture in effectively distinguishing and classifying fine art images.

Moreover, the introduction of the Non-local attention mechanism within this paper constitutes another pivotal innovation. This mechanism proves highly effective in amalgamating feature layers at various scales and extracting a more comprehensive set of feature information from the fine art images. The outcome is a substantial enhancement in classification accuracy. This mechanism optimally captures and integrates contextual information from different regions of the image, which, in turn, enhances the network's ability to discern subtle details and nuances in fine art images.

The confluence of these improvements in the MobileNetV2 architecture, along with the integration of the Non-local attention mechanism, collectively empowers the network to perform real-time and highly accurate fine art image classification and retrieval tasks. The combination of reduced complexity, improved feature reuse, and enhanced feature extraction ensures that the proposed MobileNetV2 network with the Non-local attention mechanism is exceptionally well-suited for the demanding requirements of fine art image analysis and classification.

Table 5 illustrates the network accuracy and the quantity of parameters explored through the implementation of GA. This guided network structure search using the genetic algorithm facilitates the judicious optimization of the distribution of convolutional filters with linear phase constraints within the MobileNe. Through this algorithmic approach, it becomes feasible to diminish the network's parameter count while concurrently augmenting its accuracy. This is achieved by fine-tuning the symmetrical filter array featuring linear phase constraints.

5. CONCLUSION

In this paper, we propose a convolutional neural network-based system for retrieving traditional cultural artworks and simulating and analyzing their performance. Firstly, the model parameters of the convolutional neural network are trained appropriately, then the trained network is used to extract features from fine art images, and an image feature library is built based on certain recognition requirements. Finally, a Softmax classifier is used to classify the query images and perform retrieval analysis within classes, and the corresponding retrieved results are obtained after similarity metrics. The experiments show that the image retrieval method based on the MobileNetV2 network used in this paper performs well on the traditional cultural class art image dataset built by this paper, and the classification accuracy of the network model is over 96%, and the average retrieval rate of the four selected classes is 98% when the returned image is 100. Compared with other network structures, the proposed MobileNetV2 network with Non-local attention Compared with other network structures, the MobileNetV2 network with the Non-local attention mechanism proposed in this paper has good retrieval performance and can meet the requirements related to art image retrieval, and the corresponding generalization ability also reaches a high level, which helps to strengthen students' understanding and cognition of traditional cultural artworks and effectively improves the level of art education and teaching quality of online courses in colleges and universities.

Table 5. Comparison of the accuracy of network parameters searched by GA

Network Type	Accuracy/%	Number of Parameters
Original network	86.29	3.22×10^6
MobileNet (1)	86.12	2.44×10^6
MobileNet (2)	86.31	2.69×10^6
MobileNet (3)	86.25	2.32×10^6

In future work, one direction could involve exploring techniques like model distillation, quantization, or architecture pruning to create compact yet efficient models. Moreover, investigating methods specifically tailored for reducing computational overhead while preserving accuracy would be essential. This could involve trade-offs between model complexity and performance metrics, requiring a delicate balance to achieve optimal results for resource-constrained platforms.

ACKNOWLEDGMENT

This manuscript was not funded, but I would like to thank the anonymous reviewers whose comments and suggestions helped improve this manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING AGENCY

This work received no funding.

REFERENCES

Albahli, S., Ayub, N., & Shiraz, M. (2021). Coronavirus disease (COVID-19) detection using X-ray images and enhanced DenseNet. *Applied Soft Computing*, *110*, 107645. doi:10.1016/j.asoc.2021.107645 PMID:34191925

Chen, Y., Lin, M., He, Z., Polat, K., Alhudhaif, A., & Alenezi, F. (2023). Consistency-and dependence-guided knowledge distillation for object detection in remote sensing. *Expert Systems with Applications*, *229*, 120519. doi:10.1016/j.eswa.2023.120519

Cheng, C., Hyungjoon, S., Hyun, J. C., & Zhao, Y. (2022). Pavement crack detection and classification based on fusion feature of LBP and PCA with SVM. *The International Journal of Pavement Engineering*, *23*(9), 3274–3283. doi:10.1080/10298436.2021.1888092

Feng, R. X., Li, T. H., & Dong, J. L. (2021). New media fine art education platform based on internet of things technology. *International Journal of Internet Protocol Technology*, *14*(3), 131–138. doi:10.1504/IJIPT.2021.117410

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y. J., & Fang, Z. W. (2020). Dual Attention Network for Scene Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3146-3154.

Gu, H., Liu, W. Y., Zhang, Y., & Jiang, X. Y. (2022). A novel fault diagnosis method of wind turbine bearings based on compressed sensing and AlexNet. *Measurement Science & Technology*, *33*(11), 115011. doi:10.1088/1361-6501/ac8276

Huang, Z., Wang, X., Huang, L. C., Wei, Y. C., & Liu, W. Y. (2020). CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PP*, (99), 603–612.

Hussain, M., Chen, T. H., & Hill, R. (2022). Moving toward Smart Manufacturing with an Autonomous Pallet Racking Inspection System Based on MobileNetV2. *Journal of Manufacturing and Materials Processing*, *6*(4), 75. doi:10.3390/jmmp6040075

Kasiselvanathan, M., Sangeetha, V., & Kalaiselvi, A. (2020). Palm pattern recognition using scale invariant feature transform. *International Journal of Intelligence and Sustainable Computing*, *1*(1), 44–52. doi:10.1504/IJISC.2020.104826

Li, C., Chen, Z., Jing, W., Wu, X., & Zhao, Y. (2023). A lightweight method for maize seed defects identification based on Convolutional Block Attention Module. *Frontiers in Plant Science*, *14*, 14. doi:10.3389/fpls.2023.1153226 PMID:37731985

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV),* 2999-3007.

Liu, Y., & Wei, Z. (2022). A Novel SVM Network Using HOG Feature for Prohibition Traffic Sign Recognition. *Wireless Communications and Mobile Computing*, *2022*, 6942940. doi:10.1155/2022/6942940

Pavithra, D., Nidhya, R., Vinothini, C., & Murugaiyan, M. (2023). *A Lightweight Attention based MobileNetv2 Model for Brain Tumor Segmentation and Severity of Tumor Classification using Support Vector Machine*. Academic Press.

Sun, Y. Y., Liu, S. X., Zhao, T. T., Zou, Z. H., Shen, B., Yu, Y., Zhang, S., & Zhang, H. (2021). A New Hydrogen Sensor Fault Diagnosis Method Based on Transfer Learning with LeNet-5. *Frontiers in Neurorobotics*, *15*, 664135. doi:10.3389/fnbot.2021.664135 PMID:34093159

Thepade, S. D., Dindorkar, M., Chaudhari, P., & Bang, S. (2022). Face presentation attack identification optimization with adjusting convolution blocks in VGG networks. *Intelligent Systems with Applications*, *16*, 200107. doi:10.1016/j.iswa.2022.200107

Tuna, M., & Trovalusci, P. (2022). Topology optimization of scale-dependent non-local plates. *Structural and Multidisciplinary Optimization*, *65*(9), 248. doi:10.1007/s00158-022-03351-5

Wang, C., Gong, W., Cheng, J., & Qian, Y. (2022). DBLCNN: Dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images. *Biomedical Signal Processing and Control*, *73*, 103451. doi:10.1016/j.bspc.2021.103451

Xiang, Q., Wang, X., Li, R., Zhang, G., Lai, J., & Hu, Q. (2019, October). Fruit image classification based on Mobilenetv2 with transfer learning technique. In *Proceedings of the 3rd international conference on computer science and application engineering* (pp. 1-7). doi:10.1145/3331453.3361658

Xu, X., Lin, M., Luo, X., & Xu, Z. (2023). HRST-LR: A Hessian regularization spatio-temporal low rank algorithm for traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, *24*(10), 11001–11017. doi:10.1109/TITS.2023.3279321

Xu, X., Qi, X., & Diao, X. (2020). *Reach on waste classification and identification by transfer learning and lightweight neural network*. Academic Press.

Zhao, T. T., Yi, X. L., Zeng, Z. Y., & Feng, T. (2021). MobileNet-Yolo based wildlife detection model: A case study in Yunnan Tongbiguan Nature Reserve, China. *Journal of Intelligent & Fuzzy Systems*, *41*(1), 2171–2181. doi:10.3233/JIFS-210859

Zhao, W., Su, Y. L., Hu, M. J., & Zhao, H. (2022). Hybrid ResNet based on joint basic and attention modules for long-tailed classification. *International Journal of Approximate Reasoning*, *150*, 83–97. doi:10.1016/j.ijar.2022.08.007

Zhong, M., Lin, M., & He, Z. (2023). Dynamic multi-scale topological representation for enhancing network intrusion detection. *Computers & Security*, *135*, 103516. doi:10.1016/j.cose.2023.103516

Zhu, Z., Xu, M., Bai, S., Huang, T. T., & Bai, X. (2019). Asymmetric Non-Local Neural Networks for Semantic Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* 593-602. doi:10.1109/ICCV.2019.00068