# A review of advances in image recognition models: From K-NN to Vision Transformer

**Shuhao Liu**

Gonzaga University,Spokane, Washington, WA99202, USA

sliu3@zagmail.gonzaga.edu

**Abstract.** In the digital age, visual data are everywhere. Image recognition has become a vital branch of artificial intelligence and is used in a wide range of fields. From medical diagnostics to autonomous vehicles, its impact is profound and far-reaching. This review visits the significant advancements in the field of image recognition, summarizes and analyzes the progress from traditional methods like K-Nearest Neighbors (K-NN) to newer models like Vision Transformer. This paper analyzes image recognition models like AlexNet, GoogLeNet, ResNet, and Vision Transformer, with the principles, strengths, weaknesses, and the significance of these models to deep learning fields, along with their applications in image recognition tasks. These models not only push the boundaries of image recognition but also find extensive applications across various domains. In the future, image recognition can continue to develop with larger and better datasets, increased computing power, and more advanced network structures.

**Keywords:** Computer vision, Deep network, Convolutional network, Vision transformer

## 1. Introduction

Image recognition technology experienced huge development in recent decades. Lots of new models and techniques appear, changing various fields like healthcare and autonomous driving. Traditional methods like K-NN use simple distance-based algorithms to classify images. As the scale and the complexity of the dataset grow, these conventional techniques encountered limitations in performance and efficiency.

The appearance of deep learning algorithms, particularly Convolutional Neural Networks, revolutionized the image recognition fields. The stacked structure makes it possible to extract high-level features from raw pixel data. Convolution operations make it efficient. CNN is at the forefront of image recognition research.

Since the Transformer model revolutionized the field of natural language processing field, researchers are inspired by such significant achievements, bringing it to the computer vision field. Despite the dominance of CNNs, recent breakthroughs in transformer make the visual Transformers receive considerable attention and undermined the dominance of CNNs in the CV field. It depends on self-attention mechanisms to model global dependencies across image patches. This novel approach yielded unprecedented performance gains, especially in tasks requiring long-range context understanding, such as scene understanding and image captioning.

This review mentions notable advancements in image recognition technology, from classical techniques to the latest state-of-the-art Vision Transformer models. This paper discusses the principles of different models and compare their strengths and weaknesses.

The paper begins by discussing traditional methods of image recognition, introducing K-NN and SVM algorithms, and analyzing their strengths and weaknesses. Following this, the convolutional network was introduced, which revolutionized the field of computer vision. Subsequently, several popular models such as VGGNet, GoogLeNet, and ResNet are presented to illustrate the evolution of convolutional networks. Lastly, the Vision Transformer is introduced, representing the latest advancement in deep learning research. The aim of this paper is to inspire individuals interested in computer vision and deep learning.

## 2. Traditional methods

The traditional methods of image recognition are based on classification algorithms like K-Nearest Neighbors and Support Vector Machines.

K-NN is a non-parametric learning algorithm, the classifier holds all training data. When a new data needs to be classified, it makes a decision based on the similarity of that data to all training data. To do that, it needs to calculate the similarity to all training data, which needs a lot of computation since it needs to store the entire training dataset, therefore, it needs a lot of memory to hold it.

SVM is another algorithm for classification, it first does some transformations to the training data, then maps it to a hig-dimension space then finds a hyperplane to divide the data into different categories. This algorithm is much more efficient and has a better performance compared to the K-NN algorithm.

But these algorithms need feature engineering, which means they need a handcrafted feature extractor to function, and the performance of the model highly depends on it. Based on feature extractor designed by humans also makes it cannot study higher level features and lack generalization ability. When training these models, the image usually needs to be transformed into a one-dimensional structure, which leads to the loss of special information of the image. In contrast, convolutional neural networks are more capable of handling image-related tasks.

## 3. Convolutional neural networks

Convolutional neural network is a deep learning model that uses convolution to extract features of the data. The model, usually includes stacked convolutional layers. In each convolutional layer, there are convolutional filters that slide over the input image and compute convolutions to generate feature maps. After convolutional layers, it is usually followed by a pooling layer, which progressively reduces the spatial dimensions (width and height) of the input volume, therefore reducing the number of parameters, computation in the network, and controlling overfitting.

This structure is good at utilizing the special information from the image. Besides, due to parameter sharing in convolutional operations, convolutional models have fewer parameters which makes it more efficient and easier to train. This feature also makes CNN exhibit translation invariance, meaning they can recognize the same object regardless of its position in the image. Therefore, CNN has a strong generalization ability.

The very first revolutionary convolutional network is AlexNet, it is also considered as a first convolutional deep learning model. It has five convolutional layers connected with three full connection layers. This model achieved top-1 and top-5 error rates of 37.5% and 17.0% respectively in the ImageNet LSVRC-2010 contest, [1] which is better than traditional methods like SVM at that time.

## 4. Evolution of Convolutional Networks

### *4.1. VGGNet*

VGG model is intended to improve the architecture of AlexNet and address the depth of ConvNet architecture design. Deep ConvNet can be done due to the use of small (3*3) convolution filters in all layers [2].

The configuration of VGGNet is quite different from the top-performing entries. Rather than using large receptive fields in the first convolutional layers like 11 × 11 and 7 × 7 in AlexNet, they use smaller 3 × 3 receptive fields throughout the whole net, which are convolved with the input at every pixel. It is easy to see that a stacking two or three 3 × 3 convolutional layers have an effective receptive field of 5 × 5 or 7 × 7 effective receptive field.

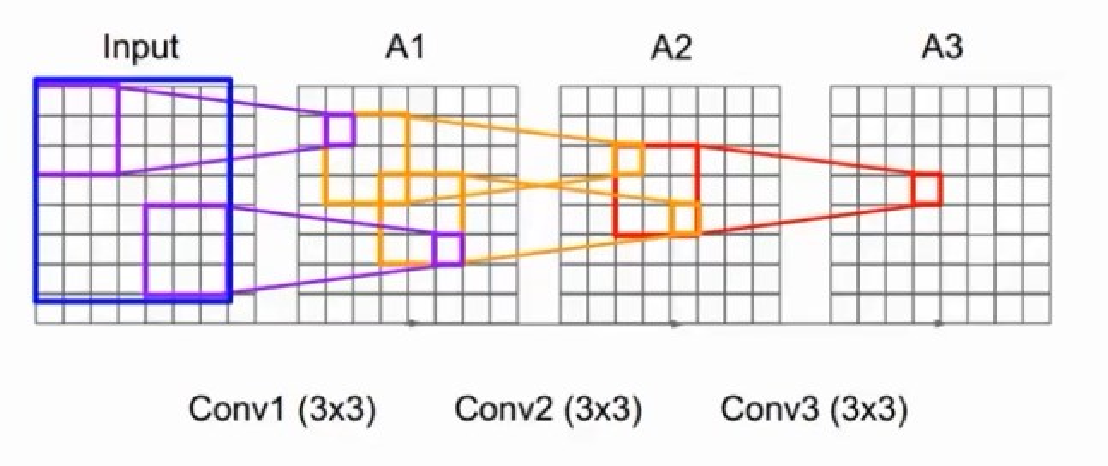

**Figure 1.** by stacking three 3 × 3 conv operations, a single pixel at the end can have a 7 × 7 receptive field, which has the same receptive field as one 7 × 7 conv operation.

By using a 3×3 convolutional filter, it has more conv layer than 7 × 7 conv, and between every two layers, there is an activation layer, which can introduce more nonlinearity to the model and make it more discriminative. Besides, it also decreases the number of parameters: a single 7 × 7 convolutional layer need 81% more parameter than a three-layer 3 × 3 convolution stack [2].

Therefore, a 3×3 convolution filter takes less computation and memory space to reach the same receptive field like 7 × 7 conv. There are a few types of VGGNet with different numbers of weight layers. The testing results of these models show that the model with the most weight layers has the best accuracy in classification tasks, which confirms the importance of depth in visual representations [2].

### *4.2. GoogLeNet*

GoogLeNet is a convolutional neural network that uses an “Inception module” to build an efficient network in a network structure model. Each inception module has the same structure, in GoogLeNet, there are stacked inception modules, inside the inception module, there are parallel convolutional and pooling operations of different filter sizes, followed by a concatenation of their outputs. Doing different convolutional operations at the same time helps the model to capture features at different scales. However, this structure presents a challenge: the number of output filters matches the number of filters in the previous layer. When merging the output of the pooling layer with the outputs of convolutional layers, the number of outputs inevitably increases at each stage. This inefficiency can cause a significant computational burden after just a few stages [3]. To overcome this problem, a 1×1 convolution is applied before the 3×3 and 5×5 convolutions, which reduces the dimension of the convolutional layer, therefore reducing the computational requirements.

Another innovation of GoogLeNet is that it does not have a full connection layer for output, it uses an average pooling layer before the output. Not using a full connection layer makes the model have fewer parameters, which makes it easier to train and reduce overfitting.

GoogLeNet is a very deep convolutional network. Therefore, ensuring effective gradient propagation through all layers was a concern. To overcome this problem, they added two auxiliary classifiers connected to these intermediate layers and expected to encourage discrimination in the lower stages in the classifier, increase the gradient signal that gets propagated back, and provide additional regularization [3]. At inference time, these auxiliary networks are discarded.

### *4.3. ResNet*

The depth of convolutional neural networks is crucially important. Deep networks naturally integrate low, mid, high-level features and classifiers in an end-to-end multilayer fashion. The richness of feature levels is enhanced by the number of stacked layers (depth) [4]. However, simply adding more layers does not necessarily improve the network. The issue of vanishing and exploding gradients prevents the model from converging, though this problem can be reduced by normalization, allowing networks with many layers to begin converging using stochastic gradient descent (SGD) with backpropagation.

Once deeper networks start converging, a degradation problem arises. As network depth increases, accuracy degrades rapidly. But this degradation is not caused by overfitting, adding more layers to a very deep model leads to higher training error.

To solve this problem, a "deep residual learning" framework was introduced. It involves adding "shortcut connections" that skip one or more layers, enabling the construction of deeper networks. This approach ensures that a deeper model should not have higher training error than its shallower counterpart.

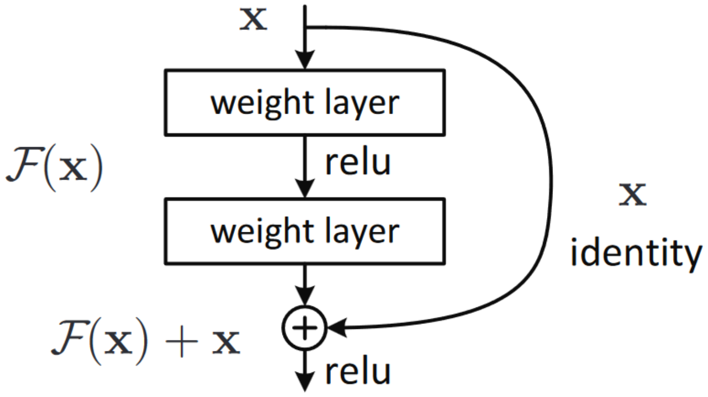
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

**Figure 2.** one building block of ResNet, demonstrates the structure of residual connection.

According to Figure 2, in ResNet, the shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers. These shortcut connections do not introduce extra parameters or computational complexity. The entire network can still be trained end-to-end using SGD with backpropagation and can be easily implemented with common libraries without needing to modify the solvers [4].

The significance of ResNet is the concept of using residual connections in deep neural networks. It makes it possible to create a deeper network. This structure can be used in not only convolutional networks, but also other networks. For example, the transformer model also uses this structure to make deeper networks.

## 5. Advanced models and techniques----the transformer model

Transformer is an attention-based encoder-decoder model. It is initially designed for natural language processing tasks, and it revolutionizes that field. Inspired by these significant achievements, pioneering work has been done on employing Transformer-like architectures in the field of computer vision (CV). Visual Transformers have demonstrated impressive performance improvements over multiple benchmarks compared to modern Convolutional Neural Networks (CNNs) [5].

Compared to Convolution Neural Networks, the Transformer has some differences. First, the Transformer does not have features of convolution, like parameter sharing and translation invariance. Second, CNN exhibits good parallelism in convolution operations, in Transformers, the attention mechanism requires pairwise comparisons between all positions in the sequence, which limits the extent of parallelization.

Vision transformer models can be divided into two categories, original transformer model and transformer combined with CNN. Since the Transformer model only accepts sequential inputs, the input image in ViT is first split into a series of non-overlapping patches, which are then projected into patch embeddings. This approach allows ViT to achieve results on multiple image recognition benchmarks that are similar to or even superior to the most prevailing CNN methods. However, its generalization capability tends to diminish with limited training data.

The convolution operation is good at processing local and shallow semantic features, while the Transformer has powerful global modeling capabilities, effectively attending to high-level semantic features. The combined CNN and Transformer model leverages the strengths of both CNNs and the original Transformer model, achieving a higher upper bound than CNNs when sufficient training data is available without complex assumptions. However, it performs poorly and converges slowly on small datasets [5].

## 6. Conclusion

In conclusion, the field of image recognition witnessed a remarkable evolution, with innovations and breakthroughs in artificial intelligence. From the conventional techniques of K-Nearest Neighbors to the forefront architectures of Vision Transformers, the journey is characterized by a pursuit of greater accuracy, efficiency, and scalability.

Starting with traditional methods in Chapter 2, it was concluded that K-NN and SVM are limited by the need for feature engineering. Convolutional Neural Networks (CNNs), as discussed in Chapter 3, revolutionized the field by enabling the extraction of high-level features from raw pixel data without extensive preprocessing. Chapter 4 focused on the evolution of CNN architectures, illustrating significant enhancements in model design that led to improvements in accuracy and efficiency. The introduction of models like VGGNet, GoogLeNet, and ResNet highlighted the importance of architectural depth and innovative design principles such as inception modules and residual connections for performance improvement. Chapter 5 introduced Vision Transformers, a breakthrough that demonstrated superior performance on complex image recognition tasks and shifted the paradigm from convolution-based to attention-based feature extraction in image recognition.

This paper provides a preliminary overview of research in computer vision and deep learning; it lacks in-depth research on these models and does not cover a wide range of different models extensively. In the future, image recognition can continue to develop with larger and better datasets, increased computing power, and more advanced network structures.

## References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[2] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Dec. 2014, doi: 10.48550/ARXIV.1409.1556.

[3] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[4] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[5] Y. Liu et al., "A Survey of Visual Transformers," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3227717.