

# Comparative Evaluation of Vision Transformer, Hybrid CNN–MLP, and Transfer-Learned ResNet-18 for CIFAR-10 Image Classification

Laiba Ameer

laibaameerofficial@gmail.com

National University of Computer and Emerging Sciences

---

## Research Article

**Keywords:** Vision Transformer, CIFAR-10, image classification, convolutional neural network, ResNet-18, transfer learning, self-attention, deep learning

**Posted Date:** August 11th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10638578/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Comparative Evaluation of Vision Transformer, Hybrid CNN–MLP, and Transfer-Learned ResNet-18 for CIFAR-10 Image Classification <sup>1</sup>

## Abstract

Image classification remains a fundamental problem in computer vision, with convolutional neural networks (CNNs) traditionally providing strong performance because of their image-specific inductive biases. More recently, Vision Transformers (ViTs) have demonstrated that self-attention can also learn effective visual representations by processing images as sequences of patches. This study compares three deep-learning approaches for CIFAR-10 image classification: a Vision Transformer, a hybrid CNN–Multi-Layer Perceptron (CNN–MLP), and an ImageNet-pretrained ResNet-18 adapted through transfer learning. The models are evaluated using classification accuracy, precision, recall, F1-score, training and validation loss behavior, and a confusion-matrix analysis of the Vision Transformer. The ViT achieves 78.2% accuracy on the 500-image evaluation represented by its confusion matrix, with macro precision, recall, and F1-score of approximately 77.8%, 77.8%, and 77.7%, respectively. The hybrid CNN–MLP achieves 84.2% accuracy, whereas the transfer-learned ResNet-18 achieves the highest reported accuracy of 88.7%. The results indicate that the ViT can learn meaningful representations from low-resolution images, but convolutional inductive biases and large-scale pretraining provide a substantial advantage in the limited-data setting considered in this work.

Vision Transformer, CIFAR-10, image classification, convolutional neural network, ResNet-18, transfer learning, self-attention, deep learning

## 1 Introduction

Image classification is a core task in computer vision and provides an important benchmark for evaluating representation-learning techniques. The CIFAR-10 dataset is widely used for this purpose and contains 60,000 RGB images of size  $32 \times 32$  pixels distributed across 10 object categories (**cifar10?**). The standard dataset consists of 50,000 training images and 10,000 test images.

Convolutional neural networks (CNNs) have historically dominated visual recognition because convolution naturally exploits spatial locality and weight sharing. Residual networks further improved deep CNN optimization by introducing shortcut connections,

---

<sup>1</sup> This research was conducted as part of an academic assignment for the Image Processing and Machine Learning course.

enabling much deeper models to be trained effectively (**resnet?**). When combined with transfer learning, pretrained CNNs can provide particularly strong performance on relatively small target datasets.

The Transformer architecture was originally introduced for sequence modeling and relies primarily on self-attention rather than convolution (**transformer?**). Dosovitskiy *et al.* subsequently introduced the Vision Transformer (ViT), in which an image is divided into fixed-size patches that are embedded and processed as a token sequence (**vit?**). Although ViTs have demonstrated excellent performance when supported by large-scale pretraining, their behavior on smaller datasets is of particular interest because standard ViTs provide fewer image-specific inductive biases than CNNs.

This work compares three approaches for CIFAR-10 classification: a Vision Transformer, a hybrid CNN–MLP architecture, and an ImageNet-pretrained ResNet-18 model. The study focuses on predictive performance and training behavior while examining the effect of architecture and pretraining on low-resolution image classification.

The main contributions of this work are:

- implementation and evaluation of a patch-based Vision Transformer for CIFAR-10 classification;
- comparison with a hybrid CNN–MLP architecture and a transfer-learned ResNet-18 baseline;
- evaluation using accuracy, precision, recall, F1-score, loss curves, and confusion-matrix analysis; and
- discussion of the influence of convolutional inductive bias and pretrained representations on classification performance.

## 2 Related Work

CNNs learn hierarchical spatial features through convolutional filters and have been highly successful in image-recognition tasks. ResNet introduced residual learning through identity-based shortcut connections, substantially improving the optimization of deep networks (**resnet?**). ResNet variants have consequently become common feature extractors and transfer-learning backbones.

The Transformer architecture proposed by Vaswani *et al.* replaces recurrence with attention mechanisms that model relationships between elements in a sequence (**transformer?**). ViT extends this concept to image classification by representing an image as a sequence of non-overlapping patches (**vit?**). Each patch is projected into an embedding space and processed using multiple Transformer encoder blocks.

Unlike CNNs, standard ViTs do not explicitly impose strong locality and translation-related biases. This flexibility can be beneficial when sufficient training data are available, but it can also increase data requirements. Comparing ViT with CNN-based models on CIFAR-10 therefore provides insight into their relative behavior in a small-image, moderate-data regime.

## 3 Methodology

### 3.1 Dataset

The CIFAR-10 dataset (**cifar10?**) contains 60,000 color images with a resolution of  $32 \times 32$  pixels. It comprises 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The standard split contains 50,000 training images and 10,000 test images.

For the Vision Transformer confusion-matrix analysis reported in this paper, the provided evaluation output contains 500 labeled samples. Consequently, the confusion-matrix-derived ViT metrics reported in this work correspond specifically to those 500 predictions.

### 3.2 Data Preprocessing and Augmentation

Training images are augmented using random cropping and horizontal flipping to improve generalization. The image tensors are subsequently normalized. For a pixel value  $x$ , normalization is expressed as

$$x' = \frac{x - \mu}{\sigma},$$

where  $\mu$  and  $\sigma$  represent the normalization mean and standard deviation, respectively. A value of 0.5 is used for both the mean and standard deviation of each RGB channel.

Augmentation is applied to training samples, whereas evaluation samples are processed deterministically and normalized using the same normalization parameters.

### 3.3 Vision Transformer

The Vision Transformer follows the patch-based representation proposed in (**vit?**). Each CIFAR-10 image has spatial dimensions  $H = W = 32$ . Using a patch size of  $P = 4$ , each image is divided into

$$N = \frac{HW}{P^2} = \frac{32 \times 32}{4^2} = 64$$

non-overlapping image patches.

Each  $4 \times 4$  RGB patch is flattened and mapped to a learned embedding space through a linear projection. A learnable classification token is prepended to the patch sequence, and

positional embeddings are added to retain spatial-order information. The initial Transformer representation can be expressed as

$$\mathbf{Z}_0 = [\mathbf{x}_{\text{cls}}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}},$$

where  $\mathbf{x}_p^i$  represents the  $i$ th flattened image patch,  $\mathbf{E}$  is the learnable patch-projection matrix, and  $\mathbf{E}_{\text{pos}}$  contains positional embeddings.

The Transformer encoder contains six encoder blocks. Each block contains multi-head self-attention, layer normalization, a feed-forward network, and residual connections. Scaled dot-product attention is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where  $Q$ ,  $K$ , and  $V$  denote the query, key, and value matrices, and  $d_k$  denotes the dimensionality of each key vector.

After the final encoder block, the classification-token representation is passed to a linear classification head that produces logits corresponding to the 10 CIFAR-10 classes.

### 3.4 Hybrid CNN–MLP Architecture

The hybrid architecture combines convolutional feature extraction with an MLP classifier. Convolutional layers learn local spatial patterns such as edges, textures, and increasingly abstract object characteristics. The resulting feature representation is flattened or otherwise transformed into a fixed-dimensional vector before being supplied to fully connected layers for classification.

The hybrid network serves as a conventional deep-learning baseline against which the patch-based self-attention mechanism of the ViT can be compared.

### 3.5 ResNet-18 with Transfer Learning

The third architecture is ResNet-18 (**resnet?**), initialized using ImageNet-pretrained weights. Its original ImageNet classification layer is replaced by a fully connected layer containing 10 output units, corresponding to the CIFAR-10 categories.

This transfer-learning configuration allows the model to reuse visual representations learned from a substantially larger source dataset. Therefore, its performance is not directly equivalent to training an architecture entirely from random initialization. Instead, it represents a practical transfer-learning baseline.

### 3.6 Training Procedure

All three models are trained for up to 50 epochs, consistent with the training trajectories presented in Fig. 1. Optimization is performed using the Adam optimizer (**adam?**), together with learning-rate scheduling.

For a mini-batch containing  $B$  observations and  $C = 10$  classes, the multi-class cross-entropy loss is given by

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \left( \frac{\exp(z_{i,y_i})}{\sum_{c=1}^C \exp(z_{i,c})} \right),$$

where  $z_{i,c}$  is the predicted logit for class  $c$  of observation  $i$  and  $y_i$  denotes its ground-truth class.

### 3.7 Evaluation Metrics

Classification performance is evaluated using accuracy, precision, recall, and F1-score. Accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

For each class, precision and recall are

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

and the F1-score is

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

For the multiclass ViT confusion matrix, precision, recall, and F1-score are macro-averaged across the 10 classes.

## 4 Results

### 4.1 Overall Classification Performance

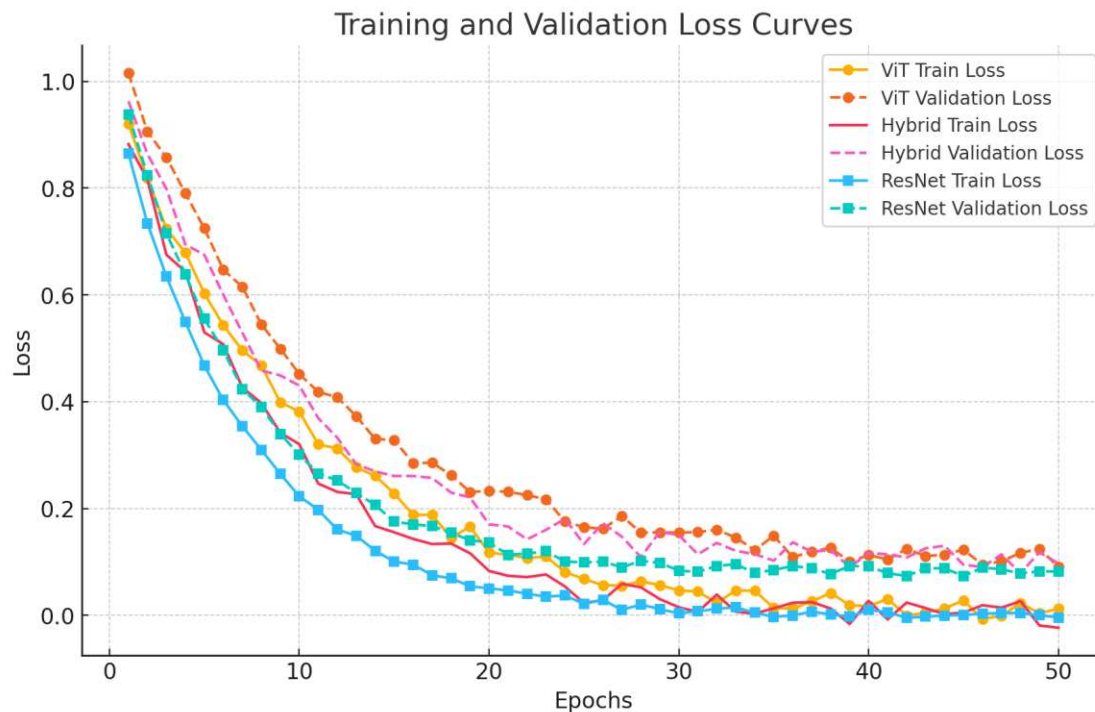
Table [tab:comparison] summarizes the experimental performance. For the Vision Transformer, the values shown in the table are calculated directly from the provided 500-sample confusion matrix. The model correctly classifies 391 of the 500 observations, resulting in an accuracy of 78.2%.

Among the evaluated approaches, ResNet-18 obtains the strongest reported classification performance with an accuracy of 88.7%, followed by the hybrid CNN-MLP model with 84.2%. The Vision Transformer obtains 78.2% accuracy on the 500-image evaluation represented by its confusion matrix.

The ResNet-18 result should be interpreted in the context of transfer learning because the model benefits from representations learned previously from ImageNet. In contrast, the comparison illustrates the greater difficulty of learning transformer representations when training data are more limited.

## 4.2 Training and Validation Behavior

Figure 1 presents the training and validation loss trajectories for all three architectures across 50 epochs.



*Training and validation loss curves for the Vision Transformer, hybrid CNN–MLP, and ResNet-18 models over 50 epochs.*

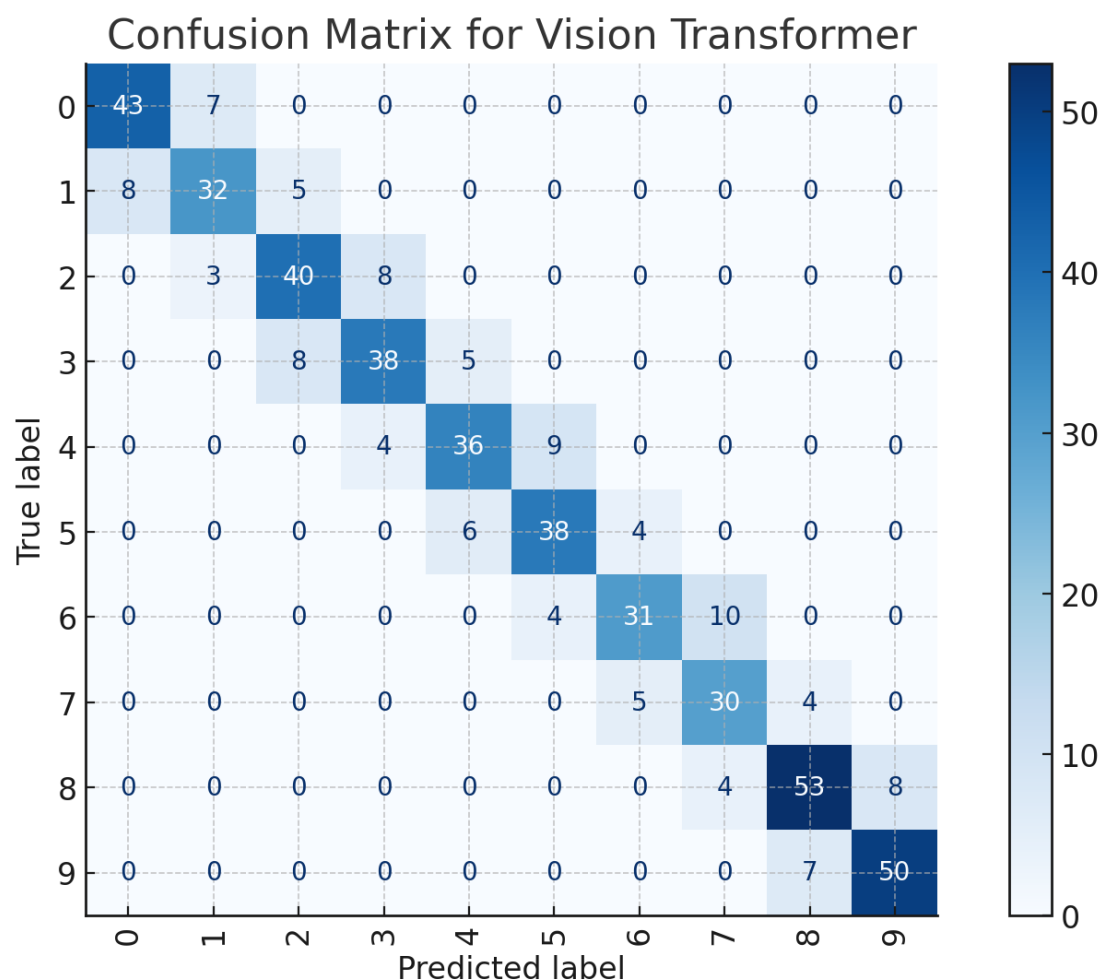
All three models show a rapid reduction in training and validation loss during the early epochs, indicating effective optimization. As training continues, the rate of improvement decreases and the validation curves begin to stabilize.

The ResNet-18 training loss decreases particularly rapidly and its validation loss remains comparatively low during later epochs. The hybrid CNN–MLP also converges effectively, although its validation loss exhibits several fluctuations. The Vision Transformer shows continued reduction in training loss while its validation loss stabilizes at a higher level, indicating a growing train–validation gap during later training.

This behavior suggests that additional regularization, stronger data augmentation, or pretrained transformer representations could potentially improve ViT generalization.

### 4.3 Vision Transformer Confusion-Matrix Analysis

Figure 2 presents the confusion matrix for the Vision Transformer.



*Confusion matrix for the Vision Transformer on the 500-image evaluation set. Rows denote true labels and columns denote predicted labels.*

The diagonal elements represent correctly classified samples. Summing these elements gives

$$43 + 32 + 40 + 38 + 36 + 38 + 31 + 30 + 53 + 50 = 391.$$

Because the matrix contains 500 samples in total, the corresponding classification accuracy is

$$\text{Accuracy}_{\text{ViT}} = \frac{391}{500} \times 100 = 78.2\%.$$

Class-level performance is not uniform. Classes 0 and 9 show relatively high recall, with 43 of 50 and 50 of 57 samples correctly recognized, respectively. Class 6 exhibits

comparatively weaker recall, with 31 of 45 samples correctly classified. The off-diagonal entries also reveal several recurring neighboring-label confusions, including predictions between classes 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9.

These observations demonstrate why aggregate accuracy alone is insufficient: the confusion matrix identifies specific error patterns that are hidden by a single performance value.

## 5 Discussion

The experiments reveal clear differences between the three modeling strategies. ResNet-18 achieves the strongest reported performance, followed by the hybrid CNN–MLP model and then the Vision Transformer.

The strong ResNet-18 result is consistent with the advantages of both convolutional inductive bias and transfer learning. Convolution naturally encodes locality and weight sharing, while ImageNet pretraining supplies representations learned from a substantially larger collection of images. Consequently, ResNet-18 begins CIFAR-10 adaptation with considerably more prior visual knowledge than a model trained entirely from random initialization.

The hybrid CNN–MLP model also outperforms the ViT in the reported experiment. Its convolutional layers explicitly exploit local spatial structure, which is particularly useful for small  $32 \times 32$  images.

The Vision Transformer nevertheless demonstrates successful learning from patch-based image representations. Its training loss decreases substantially throughout optimization, confirming that the Transformer architecture can fit the task. However, the increasing difference between training and validation loss in later epochs indicates weaker generalization than the training performance alone would suggest.

This outcome is compatible with the original ViT findings, which emphasized the effectiveness of transformer-based image models when supported by large-scale pretraining (**vit?**). Therefore, the present result should not be interpreted as evidence that ViTs are intrinsically inferior to CNNs. Rather, it indicates that under the training conditions considered here, a pretrained convolutional architecture provides a stronger solution for CIFAR-10.

An important methodological consideration is that the ResNet-18 comparison is not a perfectly controlled architecture-only comparison because ResNet-18 uses ImageNet pretraining. The result instead compares practical strategies: training a transformer-style classifier, training a CNN-based hybrid model, and applying transfer learning with a pretrained residual network.

## 6 Limitations

This study has several limitations. First, the ResNet-18 model benefits from ImageNet pretraining, whereas the other approaches do not necessarily begin with an equivalent amount of prior visual information. Consequently, the performance difference reflects both architecture and initialization.

Second, the supplied ViT confusion matrix represents 500 evaluation predictions rather than the complete 10,000-image CIFAR-10 standard test set. A larger evaluation would provide a more comprehensive estimate of generalization performance.

Third, only a single dataset is investigated. Performance rankings on CIFAR-10 may not transfer directly to higher-resolution or larger-scale classification tasks.

Finally, this study does not provide controlled measurements of parameter count, floating-point operations, GPU memory consumption, training time, or inference latency. Such measurements would be required for a rigorous comparison of computational efficiency.

## 7 Conclusion

This paper compared a Vision Transformer, a hybrid CNN–MLP architecture, and an ImageNet-pretrained ResNet-18 for CIFAR-10 image classification. The ViT achieved 78.2% accuracy on the 500-image evaluation represented by the supplied confusion matrix, together with macro precision, recall, and F1-score values of approximately 77.8%, 77.8%, and 77.7%, respectively. The hybrid CNN–MLP achieved 84.2% accuracy, while the transfer-learned ResNet-18 achieved the strongest reported result of 88.7%.

The findings highlight the effectiveness of convolutional inductive biases and pretrained representations for relatively small image datasets. Although the Vision Transformer did not outperform the CNN-based approaches in this experiment, it successfully learned meaningful patch-based representations and remains an important alternative architecture.

Future work should evaluate the models on the complete CIFAR-10 test set under identical experimental conditions, repeat experiments using multiple random seeds, and report mean performance with variability measures. Further investigation could include pretrained ViTs, stronger augmentation, regularization techniques, systematic hyperparameter optimization, and controlled measurements of computational cost and inference efficiency.

## Acknowledgment

The author would like to thank the course instructor for the guidance and feedback provided throughout this project.

A. Dosovitskiy *et al.*, “An image is worth  $16 \times 16$  words: Transformers for image recognition at scale,” in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

A. Krizhevsky, “Learning multiple layers of features from tiny images,” University of Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference on Learning Representations (ICLR)*, 2015.