*Journal of Harbin Engineering University*
*ISSN: 1006-7043*

Vol 46 No. 7
July 2025

# A Practical Exploration of Vision Transformers for Classification and Object Detection

**Aniket Dhage[1], Ayush Purandare[2], Sanika Butle[3], Aniket Patil[4], Kishan Chouhan[5], Dr. Laxmi Bewoor[6]**

[1, 2, 3, 6] Vishwakarma Institute of Information Technology, Computer Department, Pune, Maharashtra, India
[4, 5] IFM Engineering Private Limited, Pune, Maharashtra, India
Email: [1] aniket.22210283@viit.ac.in, [2] ayush.22320017@viit.ac.in, [3] sanika.22210475@viit.ac.in,
[4] Aniket.Patil@ifm.com, [5] kishan.chouhan@ifm.com, [6] laxmi.bewoor@viit.ac.in

**Abstract**

Vision Transformers (ViTs) have outperformed traditional Convolutional Neural Networks (CNNs) by integrating global context information into images with self-attention mechanisms. Thus, this study proceeds to systematically determine applications of ViTs in image classification and object detection. It methodically evaluates the efficacy of ViTs concerning tokenization, self-attention, and feedforward layers in real-world scenarios. In data augmentation, hyperparameter tuning, and proprietary datasets, various strategies were implemented that enhance model generalization. Furthermore, attention map visualizations increase interpretability and provide insights into model predictions. The study accentuates ViTs' potential for improved classification accuracy, superior object detection, and explainable AI in deep learning applications.

**Keywords**: Vision Transformers, Image Classification, Object Detection, Deep Learning, Self-Attention.

## 1. Introduction

A completely new approach for processing images is adopted by Vision Transformers (ViTs) applying Transformer's mechanisms originally designed for handling natural languages, thereby transforming the entire perspective of image processing. Generally, CNNs are useful for analyzing images, but they lack the ability to consider global context information [1]. ViTs bypass this by using patch sequences to feed the images, allowing for a construction of global dependencies without convolutional layers at all [10].

Thus, this work is to investigate the ability of ViT in object recognition and image classification based on CIFAR-10 and washer dataset. The work gives comprehensive detail of ViT design, training, evaluation of performance, and applications in real-world scenarios. They are remarkably well-suited for very common tasks like creating an object recognition pipeline for semantic segmentation or anomaly detection due to their global long-range dependency and perhaps for most common perception tasks as well [5]. New techniques may lead ViTs to a place higher than CNNs, given more data and compute [1].

The objectives of this exercise include the design of ViT Models, training them, and performance assessment with various commonly existing datasets to test their usability in day-to-day real-life applications. This result assesses the training mechanism types and the architectural configurations that enhance the success of these models.

## 2. Literature Survey

For a number of years, there was a keen interest in optimizing and establishing Vision Transformers (ViTs). A lot of research was done, enhancing their applicability in different domains. The Vision Transformer (ViT) architecture was proposed by Dosovitskiy et al. [1] and found to be useful for large-scale image datasets like ImageNet, which used pre-training on massive datasets. For instance, the transformer was originally proposed by Vaswani et al. [2] to show that the self-attention mechanism could be used in the learning of long-range dependencies, which essentially means that it emphasized the importance of modeling sequences in large arrays of data. In [3], the authors introduce self-supervised learning methods for ViTs to study emergent behaviors that improve performance without the need for human labeling. Moreover, ResNet

created by He et al. [4] provided a strong baseline for ViT models in comparison due to its effective residual connections.

Against this background, Raghu et al. [10], quoting in contrast to CNNs, state that ViTs are inherently superiors in their ability to capture global context information, while Yuan et al. [11] argued for the benefits of hybrid models incorporating both CNNs and ViTs so as to enhance efficiency, especially on small datasets. In addition, Li et al. [12] examined the effect of positional embeddings on ViT performance, accentuating their significance for benefiting image identification tasks.

Training optimizations were proposed by Touvron et al. [5] regarding attention-based distillation techniques for training data-efficient transformers and thus effectively reducing their training data requirement. Chen et al. [6] studied contrastive learning strategies, while on augmenting ViT training via data augmentation and hyperparameter tuning, Steiner et al. [7] found exposure. Grad-CAM was first introduced by Selvaraju et al. [8] to promote interpretability in ViT-based object detection tasks, thus providing insights into how attention mechanisms really work.

The scaling laws for ViTs were studied extensively by Zhai et al. [9], showing that performance improves with increases in model size and dataset volume. New strategies for token sparsification defined by Wang et al. [13] intend to lessen the computational burden, thus improving the efficiency of the ViTs. Domain adaptation techniques targeted at ViTs applied on non-natural image datasets were also studied by Sun et al. [14] to further widen the ground of applicability.

From an industrial perspective, Zhang et al. [15] pursued ViT application potential in specialized instances like washer detection, further affirming their utility in fine-grained classification tasks. All the aforementioned studies indicate the growing relevance of ViTs, leaving room for future investigations aimed at making ViTs more interpretable, computationally efficient, and adaptable to various disciplines.

## 3. Mathematical Foundations

### 3. 1. Self Attention Mechanism

Self-attention is a key component of ViTs that enables the model to assign different levels of importance to various image patches. It helps capture long-range dependencies by computing attention scores between all tokens in the input sequence [2]. The self-attention mechanism is mathematically formulated as:

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where the Query matrix (Q) is the current token to be compared. The Key matrix (K) represents the tokens that are compared with the query. The Value matrix (V) contains the actual token values. The dimension of the keys, denoted as dkd_kdk, is used to scale the dot product to stabilize training.

The softmax function normalizes attention scores to sum to one, allowing the model to focus more on significant image patches while suppressing background noise. Since self-attention operates on all tokens simultaneously, its computational complexity is, making it expensive for large images. Optimization techniques like sparse attention and low-rank factorization can help mitigate this cost.

### 3. 2. Multi Head Attention

ViTs leverage multi-head attention to learn multiple feature representations simultaneously. By applying different sets of learned projections, each attention head captures unique relationships within the input data. The multi-head attention mechanism is defined as:

$$\text{MultiHead(Q,K,V)=Concat}(head_1, head_2, \ldots head_h)\, W^O \quad (2)$$

where each attention $head_i$ is calculated as:

$$head_i = \text{A(Q}W_i^Q\text{ ,K}W_i^K\text{ ,V}W_i^V\text{ )} \quad (3)$$

Here, $W_i^Q$ , $W_i^K$ , $W_i^V$ are learned projection matrices specific to each head. A configurable hyperparameter is the number of heads (h); Using multiple heads enables the model to capture diverse aspects of an image, improving feature extraction. However, increasing the number of heads also raises computational costs, necessitating a balance between expressiveness and efficiency.

### 3. 3. Positional Encoding

Since ViTs lack inherent spatial hierarchies (unlike CNNs), positional encoding is crucial for retaining spatial information in image sequences. Positional encodings introduce order awareness by adding unique position-based vectors to each token embedding. The encoding is defined as:

$$PE_{(pos,i)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad i \text{ odd}$$

$$\text{OR} \quad \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad i \text{ even} \qquad (4)$$

where pos is Position of the token in the sequence, i is Index of the embedding dimension and $d_{model}$ is Embedding size.

This encoding ensures that tokens maintain spatial continuity, making it easier for the model to differentiate between different patches of an image, which is essential for object detection tasks [12].

## 4. Methodology

### 4. 1. Data Preparation

Three different datasets were used in this study: CIFAR-10 and a Custom Washer Dataset. The goal of the data preparation procedure was to enhance generalization and maximize the model's performance. CIFAR-10 is a dataset consisting of 60,000 32x32 pixel images divided into ten classes for the purpose of object classification. The Custom Washer Dataset employs two classes: "present" and "not present." The WIDER FACE dataset, consisting of 32,203 images with 393,703 labeled faces, was then annotated for object detection into Easy, Medium, and Hard subsets. All images were downsampled to 224x224 pixels in preparation for the input dimension of the Vision Transformer (ViT) model. The pixel values were normalized by mean and standard deviation values calculated on the ImageNet dataset. All the images were converted into PyTorch tensors so that they were usable in deep learning models. The use of augmentation techniques such as color jittering, random cropping, and horizontal-flipping would enhance the generality of the model. For further robustness, the CutMix and MixUp strategies linearly superimpose images to generate synthetic training instances.

### 4. 2. Training Strategy

The Vision Transformer (ViT) model was trained on classification tasks with pretrained ImageNet-21k weights resulting in advantage from the features learned using a large dataset. AdamW optimizer with a weight decay of 0.01 was chosen for it to ensure convergence stably. A cosine annealing scheduler was utilized for a gradual reduction of the learning rate at the epochs after being set initially at 0.0001. It used a most suitable batch size to maximize effectiveness at 64 and overfitted by including dropout (0.1) and L2 regularization. The model was held continuously under validation checks for convergence, and early stopping employed where little improvement could be realized owing to absence of further observed improvement in accuracy during the training of the ViT model for 100 epochs.

In this regard, tests were all conducted using NVIDIA A100-PCIE-40GB GPU, to speed up training as well as to manage the computational requirement in an efficient way. Tensor computation was delegated to the GPU utilizing the PyTorch CUDA backend, which significantly reduced the training time compared to CPU runs. Torch.cuda.amp was employed to support mixed precision training (FP16), allowing for faster matrix computation and reduced memory usage without compromising numerical stability. As can be seen from the training logs, the model took 1,000 steps in an average of 1.18 seconds per iteration on the CIFAR-10 dataset, which shows a huge speedup.

We used the WIDER FACE dataset that contains a variety of images having face locations labeled to train DETR from scratch for object detection tasks. As the object detection task was huge, the AdamW optimizer along with a weight decay of 0.01 was employed. A warm-up scheduler was applied to assist the convergence during initial epochs, and the learning rate was fixed as 0.00001. A smaller batch size of 16 was used because of the significant memory utilization. Twenty epochs were used to train the model, and it was periodically assessed to monitor performance indicators and guarantee successful learning.

### 4. 3. Hyperparameter Tuning

To enhance the performance of the models, hyperparameter tuning was conducted. Batch sizes of 32, 64, and 128 were considered to ensure stability and memory efficiency, and learning rates between 1e-4 and 1e-6 were investigated to find the optimal value for the Vision Transformer (ViT). To check the impact on performance and computational cost, the number of attention heads was varied from four to sixteen. For ViT, validation loss, accuracy metrics, and inference speed issues were the driving factors behind hyperparameter tuning. Multiple learning rates and batch sizes were tried for the DEtection TRansformer (DETR), focusing on finding a balance between computational feasibility and detection accuracy. For improved model performance, weightings of classification and bounding box prediction loss

functions were also altered to favor one component over the other.

## 5. System Architecture

Data ingestion, model training, testing, and deployment were some of the first steps in a systematic manner for this research. The classification pairs and detection pairs of datasets needed to have been loaded and prepared in the first step of data ingestion. The datasets were processed for having uniformity in image size and were standardized via normalization so that these could work with the Vision Transformer (ViT) model for the classification task. The WIDER facial dataset, which is annotated with facial positions, is used in object detection at various difficulty levels. The overall ViT architecture is illustrated in (Figure 1). DETR was trained from scratch on the WIDER FACE dataset for face recognition, while ViT was fine-tuned on the ImageNet pre-trained weights. The framework architecture overview is given in (Figure 2). For ViT, learning rates, batch sizes, and attention heads were cautiously tuned, while DETR was tuned for learning rates and weighing of the loss function. Attention maps were generated from ViT to facilitate interpretability to model decisions. The architecture of the DETR model is shown in (Figure 3), and the workflow of the model is explained in (Figure 4). Testing was carried out using accuracy, precision, and recall metrics to determine the performance evaluation of the models. Lastly, the deployment phase focused on forming efficient inference pipelines for real-time prediction, so that in the real-world object recognition and image classification tasks, the models could be rightfully put to application.

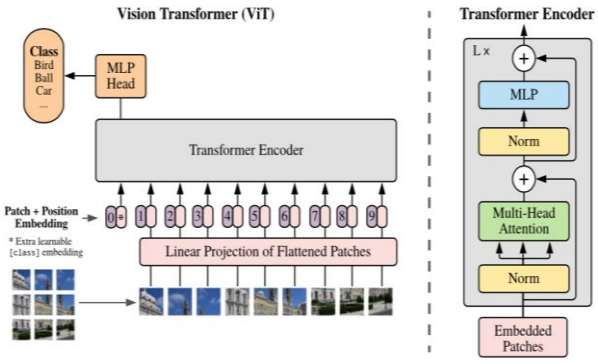


**Figure 1. Architecture of ViT**

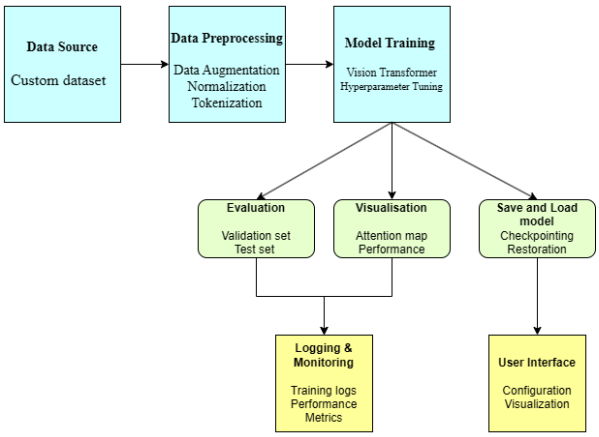


**Figure 2. System Architecture Overview**

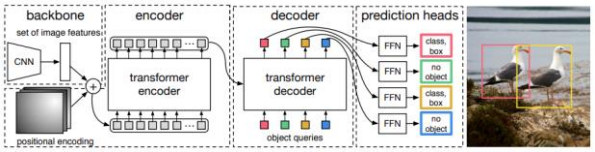


**Figure 3. DETR Model Architecture**

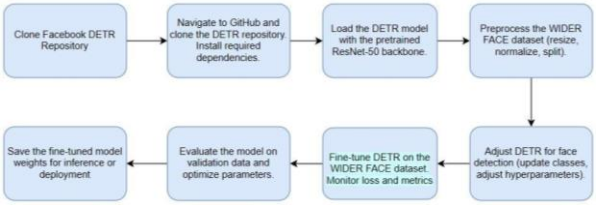


**Figure 4. DETR Model Workflow**

## 6. Results and Discussion

### 6. 1. Classification Results

The Vision Transformer (ViT) model was optimized with advanced augmentation methods and hyperparameters to achieve 98% accuracy on the CIFAR-10 dataset. The attention maps confirmed the model's interest in significant visual areas, such as petals and color patterns, which was the cause of this excellent accuracy. In Figure 5 the attention maps improved the interpretability of the model's decision-making process by graphically illustrating how it focused on these crucial locations.

The ViT model achieved 100% accuracy on the test set for the customized Washer dataset. By concentrating on crucial areas, the model successfully differentiated between the presence and absence of components, according to the attention map for this dataset. This revealed information about the decision-making process and validated the accuracy of the model's predictions (Figure 6).

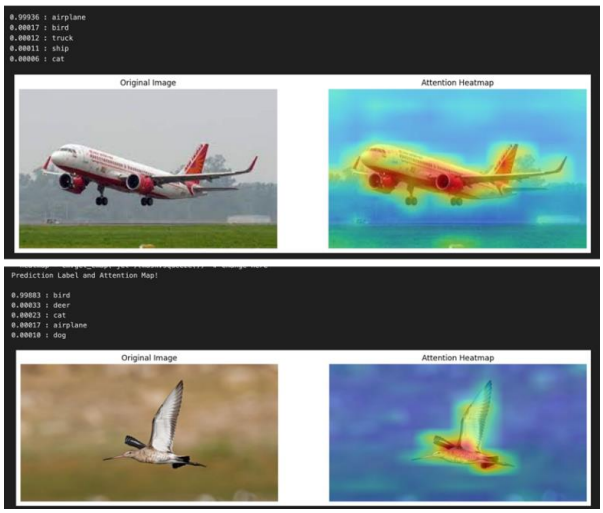


**Figure 5. Results on the CIFAR-10 dataset**

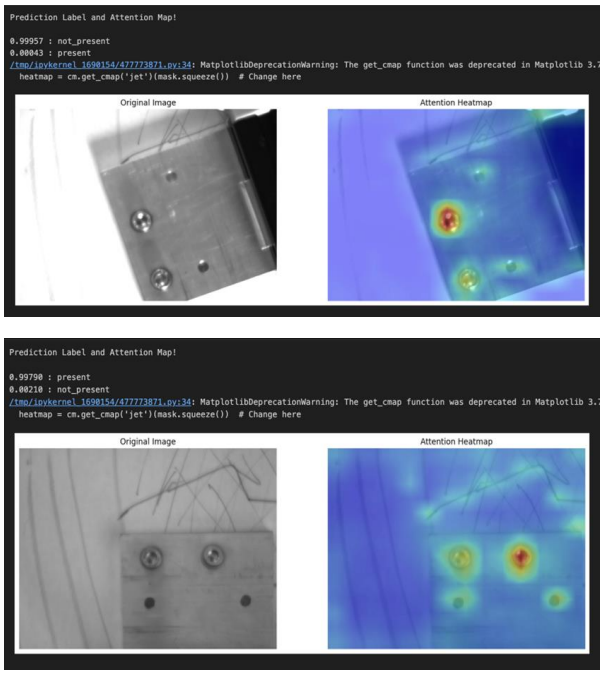


**Figure 6. Results on the Washer dataset**

### 6. 2. Object Detection Results

After being adjusted for the data, the DETR model demonstrated good precision in the object detection test using the WIDER FACE dataset. As we can see in Figure 8 Even in difficult situations, including identifying faces that are obscured or have low resolution, the model showed remarkable accuracy. The test photos' bounding box overlays confirmed the model's dependability under various circumstances.

Several loss curves are depicted throughout the training process in Figure 9, including class error, cardinality error, generalized IoU loss, bounding box loss, overall loss, and classification loss. The model's performance with trained weights is shown by the solid line, and the baseline performance is shown by the dotted line. These measurements' consistent fall suggests better item detection skills and efficient learning.

Both the classification and object recognition tasks' attention maps and bounding box representations offered insightful information about how the models made decisions. Attention maps guided the model's predictions in categorization by highlighting important regions such as object sections and color patterns. The model's accuracy in identifying faces in a variety of image situations was validated by the bounding box visualizations for object detection.


**Figure 7. Results on WIDER FACE dataset**

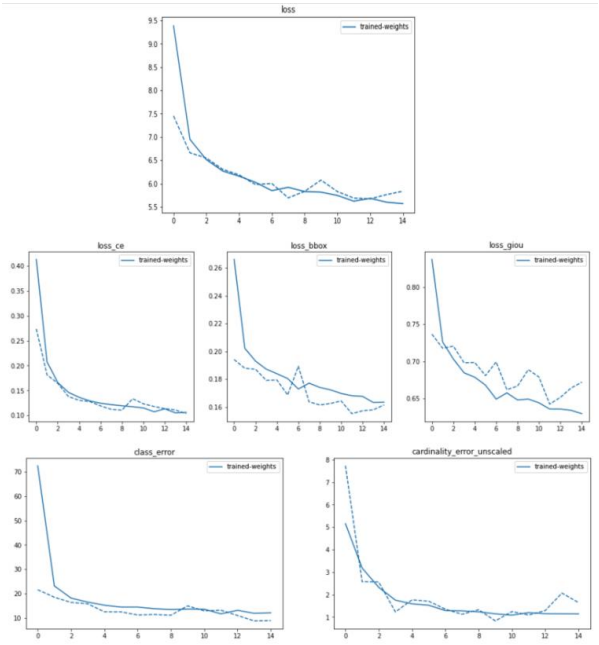


**Figure 8. Evaluation metrics for object detection**

## 7. Challenges and Learnings

A limited dataset of around 170 images posed various challenges, the most prominent being overfitting and poor generalization. Nevertheless, to enhance model generalization, data augmentation methods such as flipping, rotation, and color jitter were applied to synthetically increase the sample. It was also difficult to identify where the model was looking on the image given the complexity of interpreting attention maps from the Vision Transformer (ViT). Poor spots and regions where the model misfocused were identified by inspecting the attention maps, facilitating better model revisions. Long training times were also due to applying a short dataset and limited technology to train complicated models such as ViT and DETR. This issue was addressed through the implementation of transfer learning, using pre-trained weights, batch size and learning rate tuning, which reduced training times and improved convergence. Regularization techniques, such as dropout, weight decay, and cross-validation, were employed to minimize overfitting and improve generalization. Validation accuracy plateaued without improvement even on extending training duration, which was an indication of overfitting. Understandably, it was very hard to appreciate the reasoning behind the predictions of complex models such as ViT and DETR. To make the models more capable of being interpreted and getting more understanding in decision making, methods such as bounding boxes and attention maps were used.

## 8. Applications

Applications of Vision Transformers (ViTs) have proliferated into quite a number of fields. For instance, they assist medical image processing in detecting anomalies and segmenting tumors, thus improving the accuracy of locating pertinent areas in scans. The ViTs also perform a notable role in real-time object recognition in traffic settings for autonomous navigation, which ensures increased safety and efficiency. In industrial surveillance for factory line defects, ViTs help in detecting product faults for quality assurance. In agriculture, ViTs enable crop-facility stress monitoring by assessing aerial photos for crop health, thereby assisting the users greatly in making tricky precision agriculture decisions. Finally, ViTs serve to enhance recognition of persons in dynamic environments in the context of surveillance monitoring, thereby strengthening security measures and facilitating person identification in various settings.

## 9. Future Directions

Several avenues for future research could be undertaken by the next generation Vision Transformers (ViTs). One major direction concerns the amalgamation of hybrid models with Convolutional Neural Networks (CNNs) that integrate local-feature and global-feature recognizing architectures. This is a scenario whereby fine-grained spatial details are extracted from CNNs, while long-range dependencies are captured from ViTs through the self-attention mechanism; thus, enhancing performance in tasks that require both local and global awareness context.

Another important avenue of research is the development of further advanced ViT models, which are meant to have a sparse attention mechanism that helps in creating a better-performing ViT. The typical ViT processes its patches into a useless number for all of them, leading to exorbitant processing costs. Sparse attention techniques try to direct the model to the relevant parts of the image, thus reducing the cost without compromising on quality. This is the scalable way for making ViTs applicable in real-time applications, like video surveillance, autonomous driving, and medical image analysis, where efficiency and speed rule.

Self-supervised learning can also be a promising area of research in the case of ViTs. Several methods such as contrastive learning may enable ViTs to assimilate meaningful representations and thus rely less on large

annotated datasets. Such an outcome would be especially advantageous for domains where little labeled data is available such as satellite imagery analysis, medical diagnostics, and industrial defect detection. Overall, this may increase real-world applicability and decrease costs required since fewer human annotations are needed to train self-supervised ViTs.

Constituting yet another area of promising research is extending the capabilities of ViTs to recognize actions in videos. On the contrary, videos have information other than spatial features and time sensorially; hence, models should be developed that are efficient in capturing the movement dynamics. By applying temporal attention mechanisms, ViTs would offer better performance than CNNs and RNNs in activities such as recognizing human activity, sports analytics, and behavior monitoring. This could have far-reaching results in several areas, such as in augmenting reality, robotics, and surveillance security.

Likewise, optimization of ViTs continues to be a challenge for edge computing and mobile devices. Given the high computation cost of the transformers, research in lightweight and energy-efficient versions specifically suited to implementation on resource-constrained devices may open new avenues in mobile vision applications, wearable technologies, and the Internet of Things (IoT). In sum, though, it will have benefits for ViTs in general, even in areas such as healthcare, security, and industrial automation.

**10. Conclusion**

Through the successful extraction and learning of worldly information across an image, Vision Transformers (ViTs), by self-attention mechanisms, brought in a revolution in computer vision. ViTs are good at learning long-range relationships and contextual information to suit applications that need an understanding of the whole image, which is the opposite from the common convolutional neural networks (CNNs) that rely on local receptive fields. In this regard, ViTs were used in the present study to perform the object detection and image classification, achieving excellent results with high accuracies and thus proving promising for real-world applications. Attention maps also improved the model's focus on key regions of the image and provided relevant insight into the model's decision-making process, improving interpretability and transparency of the outcomes.

The applications of the ViT model were several in classification studies across a variety of datasets, such as CIFAR-10 and washer dataset, portraying robustness against different challenges. Robustness in attention-based mechanisms guided the model performing object detection using DETR for works such as face detection. ViTs thus behaved better than the traditional methods in accuracy and model interpretability, showing their promise in tasks that need contextual understanding and holistic feature extraction.

Numerous avenues of future research exist that are fascinating and might in fact bolster the promise of ViTs even more. An extremely relevant direction is extending ViTs to more complicated datasets and tasks, such as video action recognition, which hinges upon temporal context. ViTs could also be embedded with other neural network architectures, like convolutional backbones, to couple the advantages of local versus global feature extraction for improved performance. Strengthening ViT efficiency is another research territory, sparsity attention mechanisms being employed to reduce computational overhead and enhance model applicability in real-world situations. Generalization and the ability to learn from small datasets could be significantly increased if ViTs are trained on a small amount of labeled data, capitalizing on the strengths offered by self-supervised learning approaches like contrastive learning. All these advancements would boost the capabilities of Vision Transformers and open up pathways for a plethora of works within computer vision.

**References**

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations (ICLR).*

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is All You Need." *Advances in Neural Information Processing Systems (NeurIPS).*

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). "Emerging Properties in Self-Supervised Vision Transformers." *International Conference on Computer Vision (ICCV).*

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). "Training Data-efficient Image Transformers & Distillation through Attention." *International Conference on Machine Learning (ICML).*

[6] Chen, X., Fan, H., Girshick, R., & He, K. (2020). "Improved Baselines with Momentum Contrastive Learning." *arXiv preprint arXiv:2003.04297.*

[7] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). "How to Train Your Vision Transformer." *Advances in Neural Information Processing Systems (NeurIPS).*

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." *IEEE International Conference on Computer Vision (ICCV).*

[9] Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). "Scaling Vision Transformers." *Computer Vision and Pattern Recognition (CVPR).*

[10] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). "Do Vision Transformers See Like Convolutional Neural Networks?" *Advances in Neural Information Processing Systems (NeurIPS).*

[11] Yuan, K., Xu, G., Liu, J., Wu, T., Zhang, Y., & Li, X. (2022). "Hybrid Vision Transformer Networks for Small Dataset Applications." *Pattern Recognition Letters.*

[12] Li, C., Yang, P., Chen, L., Zhao, M., & Zhang, Q. (2023). "The Role of Positional Embeddings in Vision Transformers." *Journal of Machine Learning Research.*

[13] Wang, H., Zhao, Y., Feng, S., Liu, X., & Huang, J. (2023). "Token Sparsification in Vision Transformers for Computational Efficiency." *IEEE Transactions on Neural Networks and Learning Systems.*

[14] Sun, L., Wu, R., Zhang, H., Chen, Y., & Dong, X. (2023). "Domain Adaptation Techniques for Vision Transformers in Industrial Settings." *International Journal of Computer Vision.*

[15] Zhang, Y., Gao, M., Liu, R., & Xu, T. (2024). "Attention Mechanisms in Vision Transformers for Industrial Applications." *IEEE Access.*