ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888345

Patch Attention Excitation Based Vision Transformer for Small-Sized Datasets

Akash Verma, Shiv Ram Dubey, Satish Kumar Singh
Computer Vision and Biometrics Lab, Indian Institute of Information Technology, Allahabad
rsi2023001@iiita.ac.in, srdubey@iiita.ac.in, sk.singh@iiita.ac.in

***Abstract*—Vision Transformers have proven their mettle across a variety of computer vision problems, however, their reliance on pretraining with very large-scale datasets such as JFT-300M is also no secret, as large amounts of data is very conductive to effective feature learning. In this paper, however, we propose a novel Vision Transformer architecture PAEViT that aims to effectively learn generalized features from small-sized datasets, usually having only a few hundred images per class at most. By refining the attention scores based on patch-level interactions and modulating it to enhance the trained model's ability to focus more on task relevant patches, PAEViT is able to significantly improve upon the performance of the regular ViT model, as well as other variants of the same, when the model is being trained solely on small-sized datasets. In data-constrained situations and visual recognition tasks that do not conform well with the existing large-scale datasets, PAEViT can be used to create effective and scalable solutions with all the features and attention scores being based only on relevant data. Our code is publicly available at https://github.com/AkashVermaIN/PAEViT.**

***Index Terms*—Vision Transformers, Attention Mechanisms, Patch Attention, Data Efficiency.**

I. Introduction

The Transformer module has proven itself capable of solving both trivial and complex problems in various fields competitively and effectively [4], [15], [17], [21], and has also been used extensively and successfully on Computer Vision tasks thanks to the construction of the Vision Transformer [1], [5], [8], [10], [25]. Many variations of ViTs have also improved and built up on its advantages in different directions [6], [7], [18], [22], [26], [27], with few specifically focusing on smaller datasets [3], [9], [23], [24].

However, much of its performance still remains heavily dependent on large-scale datasets containing images in the millions [19]. Naturally, a vast majority of these images would have little to do with the task at hand, but with the traditional Vision Transformer, they remain quite necessary [5], [16], [20]. This dependence on large-scale datasets means that any model using Vision Transformers as its backbone will heavily rely on the relevant representations, however many or few they may be, in the original dataset that the model will inevitably had to have been pre-trained on [5], [16]. Given how limited such large-scale datasets are, and how difficult it is to procure and handle them, though, that leaves much of the reliability of the model entirely to chance [19]. Moreover, considering that certain specialized and sensitive computer vision problems may greatly benefit from having their models trained from relevant data only, the possibility of training vision transformer based models entirely on small-sized datasets is well worth exploring [13], [20].

Hugo Touvron et al. (2021) introduced several techniques for data efficient training of Vision Transformers, leveraging Knowledge Distillation [20]. However, their approach isn't as effective for smaller sized datasets as they are on moderately sized datasets, largely due to its reliance on distillation from a teacher model. Efficient ViT designed by Han et al. [2] also optimizes the performance of vision transformers through complexity reduction, incorporating lightweight modules to enhance training. However, while it does successfully greatly reduce computational complexity, even lightweight models can overfit or fail to generalize on smaller datasets. [2], [5]. HSViT by Chenhao Xu et al. (2022) [23] introduces a CNN based image-level feature embedding, along with enhancing the horizontal scalability of the original ViT model. However, since the most significant changes in their approach come from additional input from the images being fed through specialized convolutional layers, it deviates significantly from our goal, which is to improve the core architecture of ViTs for small datasets without relying on convolutions.

Wang et al. (2023) pursued a similar objective, with their approach being to suppress attention values that fall below a certain threshold, aptly coined Trivial Attention Weights [3]. But although the idea of filtering out these weights improves attention efficiency, reducing noise and enhancing model convergence, thus having definite merit on its own, the core attention mechanism remains the same, which may be what leads to suboptimal performance on some smaller datasets [3]. ESwin Transformer (2024) by Dazhi Yao and Yunxue Shao [24] tackles this problem by using standard shifted-window based self-attention [14] along with redesigned patch embedding and patch merging modules. However, while their implementation is able to improve on the previous performance of Swin Transformer in 300 or more epochs, other state-of-the-art ViT models designed for small datasets [3] show significant improvements on baseline ViT performance in only 100 epochs.

Meanwhile, the Shifted Patch Tokenization technique formulated by Seung Hoon Lee et al. (2021) improves Vision Transformer performance on small datasets through the preservation of spatial details, as the relationships between the shifted patches is evidently much easier to learn for the ViT model as compared to non-overlapping patches [13]. However, this technique does not modify the core Vision Transformer

Authorized licensed use limited to: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on September 24,2026 at 09:32:13 UTC from IEEE Xplore. Restrictions apply.

architecture, and can also be used on top of other solutions that may be devised to overcome the drawbacks of Vision Transformers on small-sized datasets.

The optimal solution, therefore, would be to redesign the attention mechanism in a way that allows it to learn the inter-patch relationships more efficiently without the need of pretraining on large-scale datasets. For this purpose, taking inspiration from various works in the field [2], [3], [5], [9], [13], we have designed the Patch Attention Excitation Based Vision Transformers (PAEViT) for Small-Sized Datasets.

Our key contributions are as follows:

- We introduce a novel Vision Transformer architecture called Patch Attention Excitation based Vision Transformer (PAEViT), which is tailored specifically for small-sized datasets such that the model is able to learn effectively even without pretraining.
- The Patch Attention Excitation mechanism of PAEViT is tailored to efficiently exploit the inter-patch relationships, which improves its ability to focus on task relevant patches. Since this selective scaling method is effective in improving the model's accuracy on small datasets, it can be exploited to further improve the training process for small datasets in the future.
- Through multiple experiments using varying Patch Attention Excitation factors, we are able to confirm the improved effectiveness of PAEViT compared to state-of-the-art methods on the same backbone, with the relative advantage of PAEViT only increasing as the number of input images per class is reduced, the effectiveness of our approach is validated.

II. Methodology

In this paper, we propose Patch Attention Excitation Based Vision Transformers (PAEViT) as illustrated in Fig. 1. The architecture of PAEViT significantly differs from the standard ViT model in specifically how the Multi-Head Self Attention works, with PAEViT employing additional matrices Pre-PAE and PAE, along with a Patch Attention Excitation vector $\vec{r}$ to fully exploit the inter-patch relationships that would have otherwise been missed.

The Patch Attention Excitation Module is both lightweight and efficient as it scales only based on the number of patches along with the Patch Attention Excitation factor r, thus making it independent of the complexity and size of the input data, even if it is of much higher resolution than what has been used in our experiments. It operates on input images of height H, width W, and C channels, which are divided into a non-overlapping grid of N patches post preprocessing, which includes resizing and normalization, each of which is then subsequently flattened into a linear projection of d dimension.

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C} \rightarrow \mathbf{x}_p \in \mathbb{R}^{N \times d}$$

These flattened patch vectors are then fed into a linear layer, along with the patch and positional embeddings, to generate the final input embeddings, followed by layer normalization. From this layer, our Key (K), Query (Q), and Value (V) matrices are obtained, each with the dimensions of $h \times N \times d$, where h denotes the number of attention heads being used.

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\text{Norm}(\mathbf{x}_p + \text{Embeddings}))$$

The Pre-PAE matrix $\mathbf{A}$ is then obtained by performing matrix multiplication between the previously obtained query matrix $\mathbf{Q}$ and the transposed key matrix $\mathbf{K}^T$:

$$\mathbf{A} = \mathbf{Q}\mathbf{K}^T$$

where $\mathbf{A} \in \mathbb{R}^{h \times N \times N}$. We then apply global average pooling (GAP) along the first and last dimensions of the Pre-PAE matrix $\mathbf{A}$. This Global Average Pooling (GAP) application is divided into two stages. In the first stage, GAP is applied over the last dimension (N), where we compute the mean for each head h in A.

$$\mathbf{A}_{\text{inter}}[h, i] = \frac{1}{N} \sum_{j=1}^{N} \mathbf{A}[h, i, j]$$

This reduces $\mathbf{A}$ to an intermediate Pre-PAE matrix of size $h \times N$.

$$\mathbf{A}_{\text{inter}} \in \mathbb{R}^{h \times N}$$

Then, in the second stage, GAP is applied over the first dimension, across all heads (h), reducing the intermediate Pre-PAE matrix $\mathbf{A}$ to the Pre-PAE vector $\mathbf{a}$ of N dimension.

$$\mathbf{a} = \frac{1}{h} \sum_{k=1}^{h} \mathbf{A}_{\text{inter}}[k, i], \quad \text{where} \quad \mathbf{a} \in \mathbb{R}^n$$

Within the Patch Attention Excitation module, the Pre-PAE vector $\mathbf{a}$ is then passed through a fully-connected layer, downsizing the vector to match the patch attention factor r, enabling the exploitation of the inter-patch relationship at attention score level, followed by ReLU activation to obtain the PAE vector $\vec{r} = \text{ReLU}(\mathbf{W}_1\mathbf{a} + \mathbf{b}_1)$, where $\mathbf{W}_1 \in \mathbb{R}^{r \times N}$ and $\mathbf{b}_1 \in \mathbb{R}^r$. The PAE vector $\mathbf{g}$ represents the attention modulation factors derived from the PAE module, where $\mathbf{g} \in \mathbb{R}^N$. Subsequently, another fully-connected layer followed by a sigmoid activation restructures the PAE vector to the Post-PAE vector of length N:

$$\mathbf{g} = \sigma(\mathbf{W}_2\mathbf{r} + \mathbf{b}_2)$$

where $\mathbf{W}_2 \in \mathbb{R}^{N \times r}$ and $\mathbf{b}_2 \in \mathbb{R}^N$. Sigmoid activation here ensures the modulation of the patch attention scores by constraining the value between 0 and 1. This Post-PAE vector $\mathbf{g} \in \mathbb{R}^N$ can then be duplicated across the first and last dimensions to create the Patch Attention Excitation Matrix $\mathbf{G} \in \mathbb{R}^{h \times N \times N}$, matching the dimensions of the Pre-PAE matrix A.

$$\mathbf{G}_{i,j,k} = g[j] \quad \text{for all } i = 1, \ldots, h \text{ and } k = 1, \ldots, N$$

This repetition allows the modulation of each value in the Pre-PAE matrix A by the Patch Attention Excitation vector r through the PAE matrix $\mathbf{G}$. To achieve this, the PAE matrix $\mathbf{G}$ of patch attention excitation factors undergoes element-wise

Authorized licensed use limited to: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on September 24,2026 at 09:32:13 UTC from IEEE Xplore. Restrictions apply.

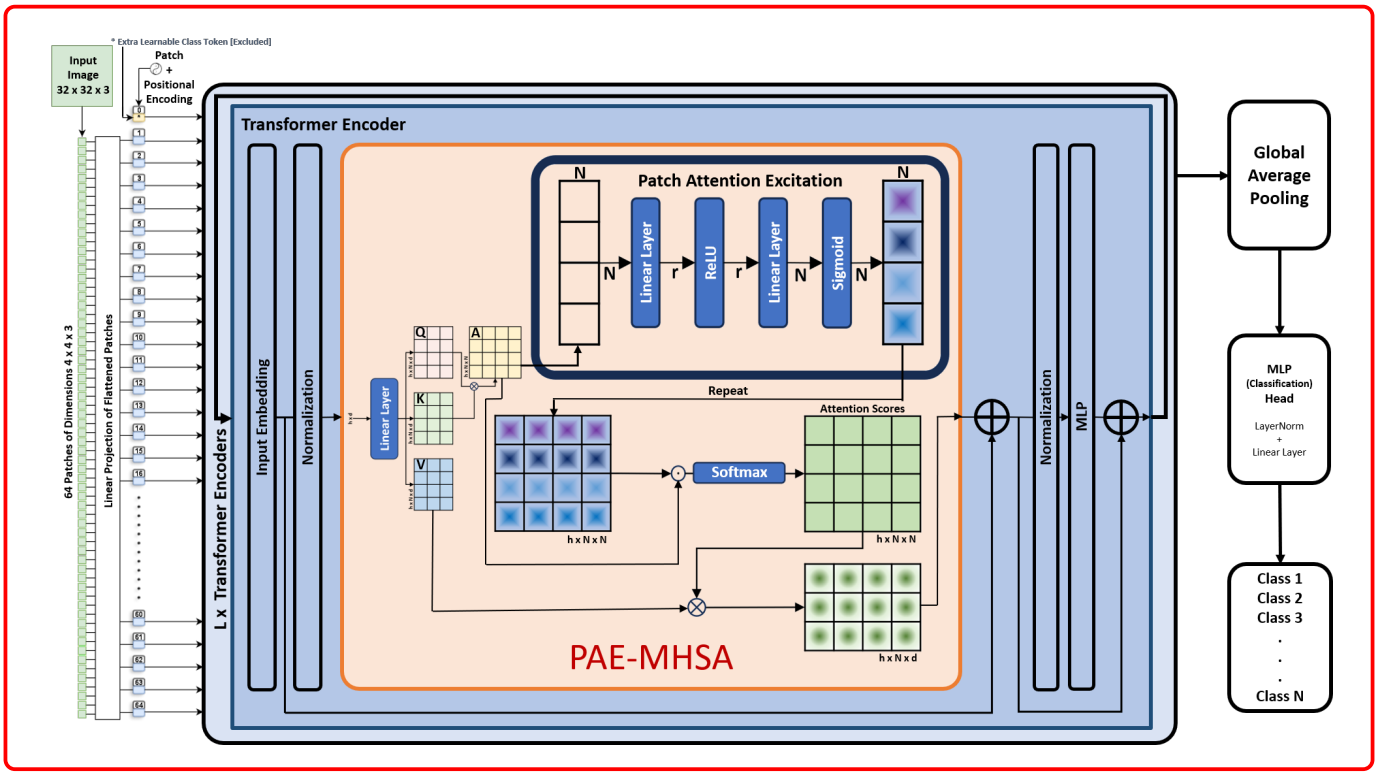


Fig. 1: Patch Attention Excitation Based Vision Transformer Architecture

multiplication with the Pre-PAE matrix $\mathbf{A}$ before the softmax operation, giving us the Post-PAE matrix $\mathbf{A}'$:

$$\mathbf{A}' = \text{softmax}(\mathbf{A} \odot \mathbf{G})$$

This Post-PAE matrix $\mathbf{A}'$ of dimension $h \times N \times N$ contains the Patch Attention Excitation enhanced attention scores, the product of which with the value matrix $\mathbf{V}$, produces the final output of the module $\hat{y}$ as follows,

$$\hat{y} = \mathbf{A}'\mathbf{V}.$$

This output $\hat{y}$ is then normalized and fed through a multi-layer perceptron, with the output becoming the input embedding of the next h minus one loops, with the final output being passed through the MLP head to obtain the class scores. The selective scaling of the raw attention scores based on the excitation factor r allows the trained model to selectively enhance or diminish the effect of the PAE module to different patches, thus incorporating the relative patch importance.

III. Experimental Setup & Results

To test the efficacy of PAEViT, we used three different datasets, namely CIFAR10 [11], CIFAR100 [11], and Tiny-ImageNet [12]. CIFAR10 dataset consists of $50,000$ images for training and $10,000$ images for testing distributed across 10 classes equally. CIFAR100 dataset contains the images of CIFAR10, but distributes the samples in 100 classes equally. TinyImageNet contains $100,000$ images distributed across 200 classes, which comes down to an average of just 500 images per category. The traditional ViT models perform very poorly on these small datasets without pretraining on extremely large datasets.

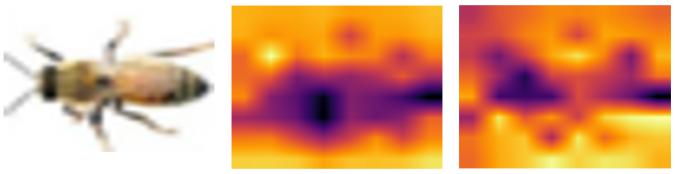
TABLE I: Classification accuracy of PAEViT compared with similar methods, with batch size 128.

Model	Epochs	Cifar10	Cifar100	TinyImageNet
ViT [5]	100	92.56	73.54	56.76
HSViT-C2A2 [23]	300	90.64	67.84	50.50
HSViT-C3A4 [23]	300	93.04	72.46	56.73
HSViT-C4A8 [23]	300	94.04	73.85	59.25
SATA-ViT [3]	100	94.18	74.93	58.77
PAEViT ($r = 4$)	100	**94.69**	76.98	61.76
PAEViT ($r = 8$)	100	94.67	77.00	61.60
PAEViT ($r = 12$)	100	94.18	**77.12**	**61.84**
PAEViT ($r = 16$)	100	94.38	77.09	61.70

The configuration being used to perform all these experiments includes the NVIDIA RTX 3090 GPU with 24 Gigabytes of VRAM, the Intel Core i9 12900K processor, and 64 Gigabytes of RAM. All the pre-existing modules and frameworks used, including PyTorch for our neural network operations, and Einops for tensor manipulations, could also be considered up-to-date as of September 2024.

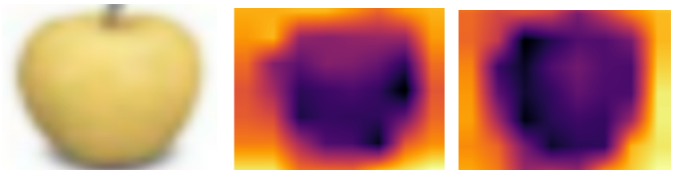
Different Patch Attention Excitation factors produce different results, however, the accuracy of models trained using PAEViT show improvement in accuracy across the board when compared with the original ViT, as well as the state-of-the-art on the same backbone by Hu et al. (2023) [3]. Among the three datasets we tested our model on, the improvement observed

Authorized licensed use limited to: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on September 24,2026 at 09:32:13 UTC from IEEE Xplore. Restrictions apply.



(a) Bee (6_2926) (b) ViT (c) PAEViT

Fig. 2: The appendages of the bee are clearer in the model trained using PAEViT.



(a) Apple (0_40607) (b) ViT (c) PAEViT

Fig. 3: The actual boundary and stem of the apple are more accurately captured by the model trained using PAEViT.

on CIFAR-10 was decent, going from 92.56% to 94.69%. In absolute terms it is a peak gain of 2.13%, and relatively our peak improvement is 2.3%, when the Patch Attention Excitation Factor r is set to 4. Compared to the state-of-the-art based on our experiments [3] the accuracy is improved by 0.51%, or a relative improvement of 0.54%. In comparison, the models trained on CIFAR-100 show even better results, with the absolute peak gain being 3.58% from 73.54% to 77.12%, relatively an improvement of 4.87% on ViTs without Patch Attention Excitation, and improving the accuracy by 2.19% compared to the state-of-the-art [3], a relative improvement of 2.92%. And on Tiny-ImageNet, the improvement is by far the most significant. The accuracy of our models is greatly improved across the board when trained with PAEViT, with the absolute peak gain being 5.08% from 56.76% to 61.84%, relatively an improvement of 8.95% on ViTs without Patch Attention Excitation, and improving the accuracy by 3.07% compared to the state-of-the-art [3], a relative improvement of 5.22%.

The test image of a bee from the CIFAR-100 dataset in Fig. 2 and that of a yellow apple from the same in Fig. 3, with the stem of the apple and the legs of the bee being faintly visible in the attention maps, and the bounding boxes also being more clear, shows how PAEViT is able to capture more details.

IV. Ablation Study

Thoroughly examining the performance of the trained model at various values of the Patch Attention Excitation factor r as shown in Table II, it can be observed that the ideal value of r varies depending on the use case, and it does not scale linearly in either extremes. When the previously ignored class token was added to PAEViT, there was a slight reduction of prediction accuracy on all three datasets. While not significant, since the pattern is observable across the board, it can be concluded that the inclusion of class token in training should be avoided for PAEViT. Next, with all other operations and modules in place except for the core Patch Attention Excitation module, the performance of the trained model was greatly reduced across the board on all the tested datasets, though it still remains slightly better than ViT. Including the class token once again into this setup shows mixed results. Running the experiment using a variety of batch sizes also gives a clear pattern, as can be seen in Table III. With the lower batch size of 64, the accuracy of most models was reduced, and the same applies to the higher batch size of 512, though the change wasn't as significant.

TABLE II: Ablation Study

Model	Cifar10	Cifar100	TinyImageNet
PAEViT ($r = 4$)	**94.69**	76.98	61.76
PAEViT ($r = 8$)	94.67	77.00	61.60
PAEViT ($r = 12$)	94.18	**77.12**	**61.84**
PAEViT ($r = 16$)	94.38	77.09	61.70
PAEViT ($r = 4$) + CLS Token	94.25	76.98	60.46
PAEViT ($r = 8$) + CLS Token	94.25	76.73	61.12
PAEViT ($r = 12$) + CLS Token	94.04	75.98	60.58
PAEViT ($r = 16$) + CLS Token	93.94	76.91	60.24
PAEViT + CLS − PAE Module	93.49	75.89	59.80
PAEViT − PAE Module	93.59	76.71	59.60

V. Conclusion

In this paper, the Patch Attention Excitation based Vision Transformer (PAEViT) model is proposed to exploit the patch relationships at attention level. A consistent improvement to performance is observed on small-scale datasets using the proposed model. PAEViT's enhanced exploitation of the inter-patch relationships at the attention score level not only make it a step above the similar existing solutions for small-sized datasets without pretraining, but can also be considered a stepping stone to build up to a truly tailored solution for certain specialized and sensitive computer vision problems where pretraining with larger and generalized datasets would otherwise be detrimental.

Model	Cifar10 (64)	Cifar10 (128)	Cifar10 (512)	Cifar100 (64)	Cifar100 (128)	Cifar100 (512)	TinyImageNet (64)	TinyImageNet (128)	TinyImageNet (512)
PAEViT ($r = 4$)	93.46	**94.69**	94.21	75.82	76.98	75.98	59.04	61.76	59.38
PAEViT ($r = 8$)	93.96	94.67	93.82	75.63	77.00	75.90	59.06	61.60	60.08
PAEViT ($r = 12$)	93.09	94.18	94.20	76.74	**77.12**	75.40	58.96	**61.84**	59.90
PAEViT ($r = 16$)	93.14	94.38	93.98	74.51	77.09	74.91	60.66	61.70	59.26
PAEViT − PAE Module	93.48	93.49	94.35	73.30	75.89	75.09	58.66	59.80	59.34
PAEViT + CLS Token − PAE Module	93.21	93.59	94.14	73.17	76.71	75.46	57.48	59.60	60.22

TABLE III: Batch Size Based Comparison of PAEViT Accuracy

Authorized licensed use limited to: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on September 24,2026 at 09:32:13 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756*, 2022.

[3] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3984–3992, January 2023.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021.

[7] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), sep 2022.

[11] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[12] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[13] Seunghoon Lee, Seunghyun Lee, and Byung Cheol Song. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 10:123212–123224, 2022.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[15] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1, 2020.

[16] Alec Radford, Jitendra Malik, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[18] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[19] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[22] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[23] Chenhao Xu, Chang-Tsun Li, Chee Peng Lim, and Douglas Creighton. Hsvit: Horizontally scalable vision transformer. *arXiv preprint arXiv:2404.05196*, 2024.

[24] Dazhi Yao and Yunxue Shao. A data efficient transformer based on swin transformer. *The Visual Computer*, 40(4):2589–2598, 2024.

[25] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[26] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. *Advances in neural information processing systems*, 34:15475–15485, 2021.

[27] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3417–3425, 2022.

Authorized licensed use limited to: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on September 24,2026 at 09:32:13 UTC from IEEE Xplore. Restrictions apply.