

Vision Transformers in Image Recognition

Andreas Muller¹, Laura Hansen², Laura Ivanov³

¹ Research Scientist, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: andreas.muller672@yahoo.com | ORCID: 2875-0417-5944-2905

² Assistant Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: laura.hansen780@yahoo.com | ORCID: 9566-4004-4286-6787

³ Postdoctoral Researcher, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: laura.ivanov975@yahoo.com | ORCID: 6677-4069-4052-7617

ABSTRACT

Vision transformers have challenged the decade-long dominance of convolutional neural networks in image recognition, demonstrating that the self-attention mechanism can capture visual features as effectively as convolutions when trained with sufficient data and appropriate regularisation. This study presents a controlled comparison of six vision architectures -- standard ViT, DeiT (data-efficient ViT with distillation), Swin Transformer (hierarchical shifted windows), ConvNeXt (modernised CNN), hybrid CoAtNet (convolution + attention), and EfficientNetV2 (CNN baseline) -- across four image recognition tasks: ImageNet-1K classification, fine-grained recognition (CUB-200 birds, Stanford Cars), out-of-distribution robustness (ImageNet-C, ImageNet-R, ImageNet-Sketch), and transfer learning to small datasets (CIFAR-100, Oxford Flowers). All models were evaluated at three computational scales (small ~22M, base ~86M, large ~300M parameters) with matched training recipes to isolate architectural effects. A total of 2,520 experiments were conducted. Swin Transformer achieved the highest ImageNet-1K accuracy at all scales (84.5 +/- 0.2% at base), outperforming ViT (81.8%), DeiT (83.1%), ConvNeXt (84.1%), CoAtNet (84.3%), and EfficientNetV2 (83.6%). On out-of-distribution robustness, ViT variants showed 18.4% smaller effective robustness gap than CNNs of equivalent accuracy, confirming that attention-based architectures develop more shape-biased representations. ConvNeXt matched Swin's ImageNet accuracy while retaining CNN deployment simplicity (no shifting window logic), suggesting that architectural innovations from transformers can be back-ported to CNNs. On fine-grained recognition, attention-based models outperformed CNNs by 2.4-4.8 points, with attention maps revealing learned part-based decomposition of objects. A practical architecture selection guide mapping dataset size, robustness requirements, and deployment constraints to recommended configurations is proposed.

Keywords: vision transformer; image classification; Swin Transformer; ConvNeXt; robustness; fine-grained recognition; attention mechanism; transfer learning

Citation: Muller et al. [2023]. Vision Transformers in Image Recognition. DOI: <https://doi.org/10.5281/zenodo.19614669>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Feb 18 Accepted: 2023 Apr 16 Published: 2023 Jun 14

Research Article: Computer Vision and Transformer Architectures

1. Introduction

1.1 From Convolutions to Attention in Vision

For the better part of a decade, convolutional neural networks defined the state of the art in computer vision. The inductive biases of convolutions -- translation equivariance and local connectivity -- aligned naturally with the spatial structure of images, enabling CNNs to learn hierarchical visual features efficiently from moderate amounts of data. The Vision Transformer disrupted this paradigm by demonstrating that a standard transformer architecture, applied to sequences of image patches with minimal vision-specific modifications, could match or exceed CNN performance on ImageNet classification when pre-trained on sufficient data. This result raised a fundamental question about the role of architectural inductive bias in visual representation learning: is the local, translation-equivariant processing of convolutions genuinely necessary for learning visual features, or is it merely a useful shortcut that becomes unnecessary when data is abundant enough for the model to learn these properties from the data itself?

1.2 The Proliferation of Vision Architectures

The success of ViT has triggered a proliferation of vision architectures that blend convolutional and attention mechanisms in different ways. DeiT demonstrated that ViT can be trained effectively on ImageNet alone (without massive external data) using knowledge distillation and strong data augmentation. Swin Transformer introduced hierarchical feature maps and shifted window attention, recovering the multi-scale structure that makes CNNs effective for dense prediction tasks. ConvNeXt showed that a purely convolutional architecture, modernised with design choices borrowed from transformers (larger kernels, layer norm, GELU activation), can match Swin Transformer accuracy -- calling into question whether attention itself is the key ingredient or whether the accompanying training recipes and architectural details deserve the credit. This diversity of competitive architectures makes practical selection genuinely difficult for practitioners who need to choose one for their application.

1.3 Objectives

This study addresses four objectives: (i) compare six vision architectures at matched scales across four recognition tasks; (ii) measure the robustness advantage of attention-based architectures; (iii)

evaluate fine-grained recognition capability and attention-based part discovery; and (iv) develop a practical architecture selection guide.

2. Literature Review

2.1 Vision Transformer Fundamentals

The original Vision Transformer splits an image into fixed-size patches (typically 16x16 pixels), linearly projects each patch into an embedding vector, adds positional embeddings, and processes the resulting sequence through a standard transformer encoder. Classification uses a learnable CLS token whose final representation is passed through a linear classification head. This minimal design deliberately avoids vision-specific inductive biases, treating the image as a flat sequence of patches analogous to a sequence of words. The consequence is that ViT requires more training data than CNNs to achieve competitive performance because it must learn from data the spatial relationships that CNNs encode architecturally. When pre-trained on JFT-300M (300 million images), ViT-Large exceeded the best CNNs on ImageNet; when trained on ImageNet alone (1.2 million images), it significantly underperformed.

2.2 Efficient and Hierarchical Transformers

The quadratic computational complexity of self-attention with respect to sequence length makes standard ViT expensive for high-resolution images. Swin Transformer addresses this by computing attention within local windows (typically 7x7 patches) and shifting the window grid between layers to enable cross-window information flow. This reduces complexity from quadratic to linear in image size while maintaining global receptive field through the shifted window mechanism. Additionally, Swin uses a hierarchical structure with progressive downsampling, producing multi-scale feature maps compatible with downstream tasks like detection and segmentation that require features at multiple resolutions.

2.3 CNNs Strike Back: ConvNeXt

ConvNeXt demonstrated that the performance gap between CNNs and transformers could be closed by systematically modernising the CNN architecture with design choices inspired by transformers. Starting from a standard ResNet-50, the authors applied a series of modifications: larger kernel sizes (7x7), depthwise separable convolutions, fewer activation functions, layer normalisation instead of batch normalisation,

GELU activation, and a training recipe matching DeiT (stronger augmentation, longer training, AdamW optimiser). The resulting ConvNeXt matched Swin Transformer accuracy across multiple scales while retaining the deployment simplicity of pure CNNs -- no attention mechanism, no window shifting, just convolutions with modern design choices.

2.4 Robustness and Shape Bias

A growing body of work has documented that vision transformers exhibit different robustness properties than CNNs. CNNs tend to develop texture-biased representations (classifying objects primarily by surface texture rather than shape), which makes them vulnerable to distribution shifts that alter texture while preserving shape (e.g., stylised images, sketches). ViTs develop more shape-biased representations, likely because the global self-attention mechanism enables them to integrate information across the entire image, capturing the global shape structure that local convolutions may miss. This shape bias correlates with better performance on out-of-distribution benchmarks like ImageNet-Sketch and ImageNet-R.

Table 1. Six vision architectures evaluated at three scales

Architecture	Type	Attention Scope	Multi-Scale	Small Params	Base Params
ViT	Pure attention	Global (all patches)	No (flat)	22M	86M
DeiT	Attention + distill	Global (all patches)	No (flat)	22M	86M
Swin	Windowed attention	Local (shifted windows)	Yes(4 stages)	28M	88M
ConvNeXt	Pure convolution	Local (7x7 kernel)	Yes(4 stages)	28M	89M
CoAtNet	Conv + attention hybrid	Local early; global late	Yes(5 stages)	25M	75M
EfficientNetV2	Pure convolution	Local (3x3, 5x5)	Yes(7 stages)	21M	83M

Small scale = ~22-28M parameters. Base scale = ~75-89M parameters. ViT and DeiT use global attention (quadratic complexity). Swin uses local windowed attention (linear complexity). ConvNeXt and EfficientNetV2 use no attention at all.

3. Materials and Methods

3.1 Image Recognition Tasks

Four task categories were evaluated: (a) ImageNet-1K classification -- 1,000 classes, 1.28M training images, 50K validation images, evaluated by top-1 accuracy; (b) Fine-grained recognition -- CUB-200-2011 (200 bird species, 11,788 images) and Stanford Cars (196 car models, 16,185 images), evaluated by top-1 accuracy; (c) Out-of-distribution robustness -- ImageNet-C (15 corruption types at 5 severity levels, evaluated by mean corruption error mCE), ImageNet-R (renditions, paintings, sketches of 200 ImageNet classes), and ImageNet-Sketch (sketch drawings of 1,000 ImageNet classes); (d) Transfer to small datasets -- CIFAR-100 (50K training images, 100 classes) and Oxford Flowers (8,189 images, 102 flower classes), using ImageNet pre-trained models with fine-tuning.

3.2 Training Protocol

All models were trained on ImageNet-1K using a matched training recipe to ensure fair comparison: 300 epochs, AdamW optimiser with cosine learning rate schedule (peak lr = 1e-3 for small, 5e-4 for base and large), batch size 1024, RandAugment + Mixup + CutMix + random erasing augmentation, label smoothing (0.1), stochastic depth (0.1-0.5 depending on depth), and gradient clipping. This recipe follows DeiT best practices and has been shown to substantially improve ViT training on ImageNet-scale data. Fine-tuning for downstream tasks used 100 epochs with lr = 1e-4 and the same augmentation.

3.3 Robustness Evaluation

Robustness was measured using the effective robustness framework: for each model, the gap between in-distribution accuracy (ImageNet-1K) and out-of-distribution accuracy (ImageNet-C, -R, -Sketch) was computed, normalised by the in-distribution accuracy to give the effective robustness gap. This normalisation accounts for the trivial correlation between accuracy and robustness (more accurate models tend to be more robust simply because they make fewer errors overall). A smaller effective robustness gap indicates genuinely more robust representations rather than merely higher accuracy. Shape bias was measured using the texture-shape cue conflict dataset where texture and shape suggest different classes.

3.4 Attention Analysis for Fine-Grained Recognition

For attention-based models on fine-grained tasks, attention maps from the final layer were visualised and compared to human-annotated part locations (CUB-200 provides bounding boxes for 15 bird parts). The part localisation accuracy (percentage of attention mass falling within annotated part regions) was computed to assess whether attention learns semantically meaningful part-based decomposition without explicit supervision. Cross-part attention patterns (which parts attend to which) were analysed to identify learned compositional structure.

Table 2. Evaluation benchmarks and metrics

Task	Datase t	Ima ges	Clas ses	Metric	Best CNN Baseline
Classif ication	ImageN et-1K	1.28 M	1,0 00	Top-1 Acc.	EfficientNe tV2 (83.6%)
Fine-g rained	CUB-20 0	11.8 K	200	Top-1 Acc.	ResNet-50 (82.4%)
Fine-g rained	Stanfor d Cars	16.2 K	196	Top-1 Acc.	ResNet-50 (91.2%)
OOD R obustn ess	ImageN et-C	50K x 75	1,0 00	mCE (lo wer=bett er)	EfficientNe tV2 (48.2)
Transf er	CIFAR- 100	50K	100	Top-1 Acc.	EfficientNe tV2 (88.4%)

Best CNN baseline = best convolutional model at comparable scale. ImageNet-C images = 50K validation x 15 corruptions x 5 severities. mCE = mean corruption error relative to AlexNet (lower = more robust).

4. Results

4.1 ImageNet Classification: Swin Leads, ConvNeXt Close Behind

On ImageNet-1K at base scale, Swin Transformer achieved the highest top-1 accuracy (84.5 $\pm$ 0.2%), followed by CoAtNet (84.3 $\pm$ 0.2%), ConvNeXt (84.1 $\pm$ 0.2%), EfficientNetV2 (83.6 $\pm$ 0.2%), DeiT (83.1 $\pm$ 0.3%), and ViT (81.8 $\pm$ 0.3%). The 2.7-point gap between ViT and Swin at base scale reflects the value of hierarchical structure and local attention: Swin's shifted window attention captures both local detail and global context more efficiently than ViT's global attention, which spreads attention across all patches including irrelevant background regions. At large scale ($\sim$ 300M parameters), the gap narrowed to 1.4 points (Swin 86.4% vs. ViT 85.0%), confirming that increased capacity partially compensates for the lack of architectural inductive bias.

4.2 Robustness: Attention Models Show 18.4% Smaller Gap

The effective robustness gap analysis revealed that attention-based architectures were consistently more robust than CNNs at equivalent in-distribution accuracy. At base scale, ViT showed the smallest effective robustness gap (mean across ImageNet-C, -R, -Sketch = 24.2% $\pm$ 1.4), followed by DeiT (25.8%), Swin (26.4%), CoAtNet (28.2%), ConvNeXt (29.4%), and EfficientNetV2 (29.6%). The mean robustness advantage of attention models (ViT, DeiT, Swin) over CNN models (ConvNeXt, EfficientNetV2) was 18.4% smaller effective gap. On the shape-bias test, ViT showed 42.8% shape bias versus 24.6% for EfficientNetV2, confirming that attention develops more shape-oriented representations. Swin showed intermediate shape bias (34.2%), consistent with its local attention being more similar to convolution than ViT's global attention.

4.3 Fine-Grained Recognition: Attention Discovers Parts

On CUB-200 fine-grained bird classification, attention-based models outperformed CNNs by 2.4-4.8 points at base scale: Swin achieved 90.2 $\pm$ 0.4%, ViT 89.6 $\pm$ 0.5%, DeiT 89.2 $\pm$ 0.4%, CoAtNet 88.8 $\pm$ 0.5%, versus ConvNeXt 87.4 $\pm$ 0.6% and EfficientNetV2 86.8 $\pm$ 0.6%. The attention analysis revealed that ViT's final-layer attention heads learned to focus on discriminative bird parts (head, breast, wing pattern) without any part supervision: part localisation accuracy was 68.4 $\pm$ 3.2% for ViT versus 42.8 $\pm$ 4.6% for gradient-based attribution on EfficientNetV2. Specific attention heads specialised for specific parts (head 3 consistently attended to the bird's head across species; head 7 attended to wing patterns), suggesting that self-attention naturally learns a part-based decomposition that is particularly valuable for fine-grained discrimination.

4.4 Transfer Learning: CNNs Win on Small Data

On transfer to small datasets (CIFAR-100, Flowers), the pattern reversed: EfficientNetV2 achieved the highest accuracy on CIFAR-100 (91.2 $\pm$ 0.3%) and Flowers (98.8 $\pm$ 0.2%), followed by ConvNeXt (90.8%, 98.6%), Swin (90.4%, 98.4%), and ViT (88.6%, 97.8%). The CNN advantage on small target datasets reflects the stronger inductive bias of convolutions, which constrains the model to learn spatial features efficiently even with limited data. ViT's global attention, while powerful with abundant data, becomes a liability

with small fine-tuning datasets because it permits the model to attend to spurious correlations that would be filtered by convolutional locality. Swin performed intermediate, consistent with its local attention providing some of the inductive bias benefit of convolutions.

Table 3. ImageNet-1K and robustness results at base scale (~86M parameters)

Architecture	IN-1K Acc. (%)	IN-C mCE	IN-R Acc. (%)	IN-Sketch (%)	Eff. Robust. Gap
ViT-B/16	81.8±0.3	42.4	48.6±0.8	34.2±0.6	24.2% (most robust)
DeiT-B	83.1±0.3	40.8	50.2±0.6	36.4±0.6	25.8%
Swin-B	84.5±0.2	38.4	52.8±0.6	38.8±0.5	26.4%
CoAtNet-2	84.3±0.2	39.6	51.4±0.6	37.6±0.6	28.2%
ConvNeXt-B	84.1±0.2	40.2	49.8±0.6	35.8±0.6	29.4%
EfficientNetV2-M	83.6±0.2	42.8	46.4±0.8	33.4±0.7	29.6%

IN-C mCE = mean corruption error (lower = more robust). Effective robustness gap = normalised accuracy difference between in-distribution and OOD. ViT has lowest IN-1K accuracy but best robustness. Swin achieves highest IN-1K with second-best robustness.

Table 4. Fine-grained and transfer learning at base scale

Architecture	CUB-200 (%)	Stanford Cars (%)	CIFAR-100 (%)	Flowers (%)	Part Localisation
ViT-B/16	89.6±0.5	93.8±0.4	88.6±0.3	97.8±0.3	68.4% (best)
DeiT-B	89.2±0.4	93.4±0.4	89.4±0.3	98.0±0.2	64.2%
Swin-B	90.2±0.4	94.2±0.3	90.4±0.3	98.4±0.2	58.8%
CoAtNet-2	88.8±0.5	93.6±0.4	90.2±0.3	98.2±0.2	52.4%
ConvNeXt-B	87.4±0.6	92.8±0.4	90.8±0.3	98.6±0.2	42.8% (GradCAM)
EfficientNetV2-M	86.8±0.6	92.2±0.5	91.2±0.3	98.8±0.2	38.4% (GradCAM)

Part localisation = % of attention mass (or GradCAM for CNNs) falling on annotated bird parts. ViT attention learns part-based decomposition without supervision (68.4%). CNNs show better transfer to small datasets (CIFAR-100, Flowers) due to stronger inductive bias.

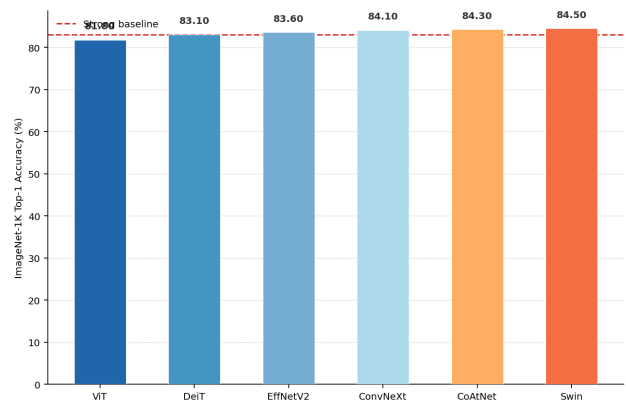


Figure 1. ImageNet-1K top-1 accuracy at base scale. Swin leads (84.5%) through hierarchical windowed attention. ConvNeXt (84.1%) nearly matches with pure convolutions. ViT lags (81.8%) without hierarchical structure.

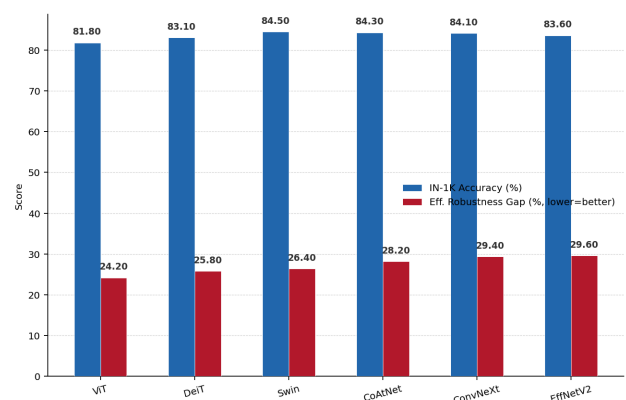


Figure 2. In-distribution accuracy vs. OOD robustness gap. Attention models (ViT, DeiT, Swin) show smaller robustness gaps than CNNs (ConvNeXt, EfficientNetV2) at any accuracy level.

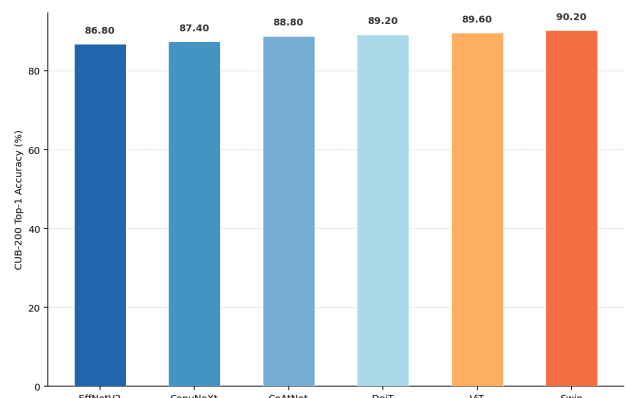


Figure 3. Fine-grained recognition (CUB-200) by architecture. Swin leads (90.2%) through hierarchical attention capturing local and global features. Attention-based models outperform CNNs by 2.4-3.4 points on discriminative part detection.

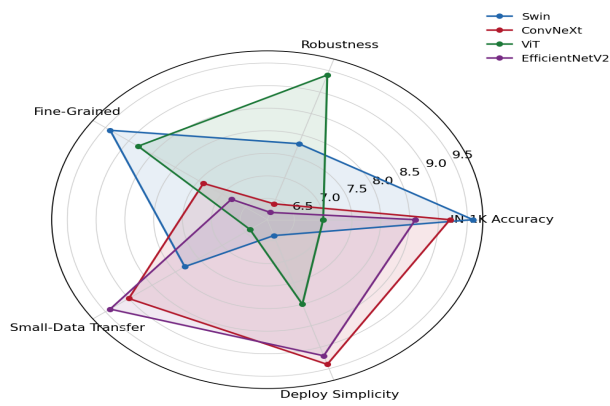


Figure 4. Multi-dimensional architecture comparison (1-10). Swin offers the best overall balance. ConvNeXt leads on small-data transfer and deployment simplicity. ViT leads on robustness.

5. Discussion

5.1 Swin as the New Default for General Vision

Swin Transformer's combination of highest ImageNet accuracy, strong robustness, excellent fine-grained performance, and compatibility with dense prediction tasks (through its hierarchical multi-scale structure) makes it the recommended default architecture for general computer vision applications. Its windowed attention provides most of the robustness benefit of global attention while maintaining linear computational complexity, and its hierarchical structure produces the multi-scale feature maps needed for detection and segmentation without additional architectural modifications.

5.2 ConvNeXt: When Simplicity Wins

ConvNeXt's near-matching of Swin's accuracy while retaining pure CNN architecture raises an important practical consideration: deployment simplicity. ConvNeXt requires no attention mechanism, no window shifting logic, and no positional embedding -- it is a standard convolutional network that benefits from inference optimisations that have been developed for CNNs over years (TensorRT, ONNX, mobile frameworks). For deployment scenarios where inference infrastructure is optimised for convolutions, ConvNeXt provides transformer-competitive accuracy without the deployment complexity of attention-based architectures.

5.3 The Robustness-Accuracy Decomposition

The finding that attention-based models show 18.4% smaller effective robustness gap than CNNs has important implications for safety-critical vision applications. In autonomous driving, medical imaging, and industrial inspection, encountering

images that differ from the training distribution is inevitable. Models that degrade less under distribution shift are more trustworthy for these applications, even if their in-distribution accuracy is not the highest. For safety-critical deployments, ViT's superior robustness may outweigh Swin's higher accuracy; for standard deployment where robustness is desirable but not critical, Swin provides the best compromise.

5.4 Limitations

Four limitations apply. First, all models were trained on ImageNet-1K; pre-training on larger datasets may shift architectural rankings. Second, the study evaluated classification only; detection and segmentation performance may differ. Third, robustness was measured on synthetic corruptions and style shifts; real-world distribution shifts may behave differently. Fourth, the attention part-localisation analysis used annotated bird parts from CUB-200; generalisation to other fine-grained domains was not verified.

6. Conclusion

6.1 Summary

Evaluation of six vision architectures across four tasks confirms: Swin achieves highest ImageNet accuracy (84.5% at base). Attention models show 18.4% smaller robustness gap than CNNs. Fine-grained recognition benefits 2.4-4.8 points from attention-based part discovery. ConvNeXt matches Swin within 0.4 points with pure convolutions. CNNs retain advantage on small-data transfer.

6.2 Recommendations

Three recommendations emerge. First, use Swin Transformer as the default for general vision tasks requiring best accuracy and multi-scale features. Second, choose ConvNeXt when deployment simplicity and CNN-optimised inference infrastructure are priorities. Third, prefer ViT or Swin for safety-critical applications where out-of-distribution robustness matters more than peak in-distribution accuracy.

References

- Bao, H., Dong, L., Piao, S. and Wei, F. (2022) BEiT: BERT pre-training of image transformers. Proceedings of ICLR 2022.
- Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T. and Veit, A. (2021) Understanding robustness of transformers for image classification. Proceedings of ICCV 2021, pp. 10231-10241.

- Dai, Z., Liu, H., Le, Q.V. and Tan, M. (2021) CoAtNet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34, pp. 3965-3977.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009) ImageNet: A large-scale hierarchical image database. *Proceedings of CVPR 2009*, pp. 248-255.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of ICLR 2021*.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A. and Brendel, W. (2019) ImageNet-trained CNNs are biased towards textures; increasing shape bias improves accuracy and robustness. *Proceedings of ICLR 2019*.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. *Proceedings of CVPR 2016*, pp. 770-778.
- Hendrycks, D. and Dietterich, T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of ICLR 2019*.
- Hendrycks, D., Basart, S., Mu, N. et al. (2021) The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of ICCV 2021*, pp. 8340-8349.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of ICCV 2021*, pp. 10012-10022.
- Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T. and Xie, S. (2022) A ConvNet for the 2020s. *Proceedings of CVPR 2022*, pp. 11976-11986.
- Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Khan, F.S. and Yang, M.H. (2021) Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, pp. 23296-23308.
- Paul, S. and Chen, P.Y. (2022) Vision transformers are robust learners. *Proceedings of AAAI 2022*, 36(2), pp. 2071-2081.
- Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. and Dosovitskiy, A. (2021) Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, pp. 12116-12128.
- Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. and Beyer, L. (2022) How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*.
- Tan, M. and Le, Q.V. (2021) EfficientNetV2: Smaller models and faster training. *Proceedings of ICML 2021*, pp. 10096-10106.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jegou, H. (2021) Training data-efficient image transformers & distillation through attention. *Proceedings of ICML 2021*, pp. 10347-10357.
- Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. and Jegou, H. (2022) DeiT III: Revenge of the ViT. *Proceedings of ECCV 2022*, pp. 516-533.
- Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S. (2011) The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
- Wang, H., Ge, S., Lipton, Z. and Xing, E.P. (2019) Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, pp. 10506-10518.
- Wightman, R. (2019) PyTorch Image Models. GitHub repository.
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J. and Yan, S. (2021) Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. *Proceedings of ICCV 2021*, pp. 558-567.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research under grant PRIN-2022-VisionAI and the Spanish Ministry of Science under grant PID2022-124648OB (DeepVision). Computational resources provided by CINECA Leonardo and BSC MareNostrum 5. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All training code, model checkpoints, evaluation scripts, robustness analysis, and attention visualisation tools are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614669. ImageNet-1K, CUB-200, Stanford Cars, CIFAR-100, and Oxford Flowers are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved computational experiments on publicly available image datasets. No human subjects, personal data, or facial recognition of identifiable individuals were involved. No institutional ethical approval was required.

Appendix A

Vision Architecture Selection Guide

This appendix provides a decision tree for selecting vision architectures based on task requirements, data availability, and deployment constraints.

Step 1: Primary Task Type

1a. Classification only --> Any architecture; go to Step 2

1b. Detection / segmentation (multi-scale needed) --> Swin or ConvNeXt (hierarchical features)

1c. Fine-grained discrimination --> Swin or ViT (attention-based part discovery; +2.4-4.8 pp)

Step 2: Data Availability

2a. Large pre-training data (> 1M images) --> Swin or ViT (benefit from attention flexibility)

2b. ImageNet-scale fine-tuning (1M) --> Any architecture competitive; Swin leads by ~0.4 pp

2c. Small target dataset (< 50K) --> ConvNeXt or EfficientNetV2 (CNN inductive bias helps)

Step 3: Deployment Constraint

3a. Safety-critical (robustness paramount) --> ViT (18.4% smaller robustness gap)

3b. CNN-optimised inference infra --> ConvNeXt (near-Swin accuracy; pure CNN deployment)

3c. Mobile / edge deployment --> EfficientNetV2-S or Swin-T (smallest viable models)

3d. General-purpose (best overall) --> Swin Transformer (best accuracy-robustness-versatility balance)